Statistics and Its Interface Volume 5 (2012) 21–37

# Protein identification problem from a Bayesian point of view

Yong Fuga Li, Randy J. Arnold, Predrag Radivojac and Haixu Tang*

We present a generic Bayesian framework for the peptide and protein identification in proteomics, and provide a unified interpretation for the database searching and the *de novo* peptide sequencing approaches that are used in peptide identification. We describe several probabilistic graphical models and a variety of prior distributions that can be incorporated into the Bayesian framework to model different types of prior information, such as the known protein sequences, the known protein abundances, the peptide precursor masses, the estimated peptide retention time and the peptide detectabilities. Various applications of the Bayesian framework are discussed theoretically, including its application to the identification of peptides containing mutations and post-translational modifications.

AMS 2000 subject classifications: 60K35.
Keywords and phrases: Shotgun proteomics, Protein identification, Mass spectrometry, Bayesian methods.

## 1. INTRODUCTION

The rapid advancement of protein analytical technologies in the past decades, in particular the instrumental development of mass spectrometry, has enabled the routine analysis of complex proteome samples at high throughput and sensitivity. The most commonly used proteomics protocol (often referred to as the *bottom-up* or *shotgun* proteomics), starts from the trypsin digestion of a complex protein mixture, followed by the analysis of tryptic peptides using liquid chromatography tandem mass spectrometry (LC-MS/MS). A shotgun proteomics dataset may consist of 10–100 thousands of tandem mass (MS/MS) spectra, each of which is generated from the fragmentation (e.g. by using collision-induced dissociation, or CID) of a peptide ion, and thus contains the information of a peptide sequence. The primary task in shotgun proteomics, known as the *protein identification problem*, is to identify the proteins in a proteome sample based on the set of MS/MS spectra acquired from the sample. This problem is usually approached by a database searching strategy, in which we attempt to infer, for each protein in a given protein database of the organism from which the proteome sample is obtained (e.g. the human), if it is present in the sample. The database searching procedure consists of two steps: 1) the *peptide identification*, in which the tryptic peptides of the proteins in the database are identified (or scored) based on their MS/MS spectra; and 2) the *protein inference*, in which the presence or absence of proteins are inferred from the identified peptides. A notable challenge of the protein inference problem is that many tryptic peptides are shared by two or more proteins from the same organism (called the *degenerate peptides*), in particular for the higher organisms like mammals [27]; and thus, a protein inference algorithm should be able to distinguish the proteins sharing degenerate peptides. Many software tools were developed for peptide identification, known as the *peptide search engines*, including SEQUEST [40], MASCOT [28], X!Tandem [8], OMSSA [16], and InSpecT [35], whereas post-processing algorithms are also available for protein inference, addressing the challenge of degenerate peptides, including the ones based on parsimony formulation such as IDPicker [41, 25] and MMP [2], and the ones based on probabilistic models such as ProteinProphet [26], MSBayesPro [21, 22], and OBM [30].

Statistical modeling methods are commonly applied to protein identification, because an LC-MS/MS proteomics experiment is a complex stochastic process. In particular, in the peptide identification step, in addition to reporting a peptide that is most likely to produce a given MS/MS spectrum (i.e. a *peptide-spectrum matching*, or PSM [13]), a peptide search engine should also provide a score that can be used to distinguish the correct PSMs from the incorrect ones. This measure can be either completely heuristic [40, 28, 8], or can be based on a statistical model of a heuristic score and thus has a probabilistic interpretation [16, 19]. A commonly used statistical measure is the *false discovery rate* (FDR), computed by using a target-decoy search strategy [11]. However, it is often not clear that these statistical measures of PSMs can be used for protein inference, i.e. to compute the likelihood of a protein being present in the sample. For instance, the FDR of an identified protein cannot be easily derived from the FDRs of the identified peptides from this protein; as a result, independent procedures are needed to derive the FDR of protein identification based on peptide search results [39, 29].

*Corresponding author.

One of the probabilistic approaches that integrates the peptide identification and protein inference steps was presented in the PeptideProphet/ProteinProphet algorithms [19, 26]. PeptideProphet used a model-based method to compute a probability for each PSM to be correct (or incorrect), and these probabilities were then used by ProteinProphet to compute a probability of each protein being correctly identified. Notably, the probabilities of PSMs corresponding to degenerate peptides were considered in the parsimonious model of ProteinProphet, and as a result, the protein probabilities from ProteinProphet can, to some extent, distinguish proteins sharing degenerate peptides [26].

An important feature of the proteomics analysis that most current probabilistic models for protein identification do not consider is that, based on prior knowledge, the proteins (and peptides) are not equally likely to be observed in a proteomics (LC-MS/MS) experiment. For example, some peptides are more likely identified than the others in an LC-MS/MS experiment because they are either more easily ionized or better fragmented owning to their physicochemical properties. It was shown that the probabilities of peptides being observed in an LC-MS/MS experiment, defined as the (standard) *peptide detectabilities*, can be predicted solely from the peptide sequences [34, 23], and can be incorporated into a Bayesian model to improve protein inference [22]. The other prior knowledge that can be used to improve protein inference include the confidence of proteins in the database (e.g. if proteins are computationally predicted from genomic sequences) and probabilities of proteins being post-translationally modified and being processed in the cell. In fact, the incorporation of the prior knowledge into protein identification is not new. The database searching approach itself is based on the assumption that only the proteins in the database can be present in the sample. In comparison, the *de novo* peptide sequencing algorithms (such as PEAKS [24], pepNovo [14] and pepHMM [37]) attempt to reconstruct the peptide sequence directly from an MS/MS spectrum. Assuming all peptides (not only peptides in a database) are equally likely to be identified, the p-value for a PSM to be derived from *de novo* sequencing by chance can be estimated by using a dynamic programming algorithm [20]. Previous studies showed that although the most likely peptide reported by the *de novo* sequencing algorithm for an MS/MS spectrum was often incorrect, correct peptide was usually the peptide in the database that receives the highest score in *de novo* sequencing [14, 24], indicating that the incorporation of the protein database as prior knowledge significantly improves protein identification.

In this paper, we describe a Bayesian framework that can be used to incorporate various prior information into peptide and protein identification, extended from our previous Bayesian model for protein inference [21, 22]. The rest of the paper is organized as follows. We will first introduce a general Bayesian framework to model the LC-MS/MS experiment and its application to incorporating various kinds of prior information in peptide identification. We will then provide a unified interpretation of *de novo* peptide sequencing and database searching algorithms based on the Bayesian framework. Next, we will show that the Bayesian model can be further extended to the protein inference step, and discuss the potential advantage of using the Bayesian framework to integrate the peptide identification and protein inference steps. Finally, we will discuss some practical issues on the implementation of the Bayesian models, including the computational complexity and the selection of appropriate prior distributions.

## 2. A GENERAL BAYESIAN MODEL FOR LC-MS/MS ANALYSES

An LC-MS/MS proteomics experiment is a complex stochastic process. The same proteome sample analyzed multiple times following the same experimental protocol on the same platform may result in different mass spectral data, and consequently lead to merely 30%–60% overlaps in peptide identifications [33, 15]. The overlap can even be $< 10\%$ when different experimental protocols (e.g. for proteolysis) or platforms (e.g. mass spectrometers) are used [15]. Taking the stochasticity of LC-MS/MS analyses into consideration, we denote the probability of observing the MS data $M$ from sample $P$ using an experimental protocol $\mathcal{T}$, as $\mathtt{P}(M|P,\mathcal{T})$. The goal of the *protein identification* can then be formulated as the problem of approximating the posterior probability

$$\mathtt{P}(P|M,\mathcal{T}) = \frac{\mathtt{P}(M|P,\mathcal{T}) \cdot \mathtt{P}(P|\mathcal{T})}{\mathtt{P}(M|\mathcal{T})}, \tag{2.1}$$

for the protein sample $P$. Assuming that the proteome sample and the LC-MS/MS analysis are independent, we have $\mathtt{P}(P|M,\mathcal{T}) \propto \mathtt{P}(M|P,\mathcal{T}) \cdot \mathtt{P}(P)$, as represented in the graphical model in Figure 3.1a. Furthermore, we neglect the MS scans in the observed MS data and only model the MS/MS spectra (denoted as $MS2$). Hence, $\mathtt{P}(P|MS2,\mathcal{T}) \propto \mathtt{P}(MS2|P,\mathcal{T}) \cdot \mathtt{P}(P)$. For the *peptide identification* step, as represented in Figure 3.1d, we consider each protein as a set of (tryptic) peptides, and thus the goal is to determine the peptide(s) that are fragmented in each MS/MS spectrum. Let $Y = (y_k)_{k=1,\dots,K}$ be the indicator matrix for peptides corresponding to the set of fragmentation spectra (each vector $y_k \in \{0,1\}^J$ for peptides corresponding to a spectrum $s_k$, where $J = |PD|$ is the size of the database), and let $PD$ (or $DB$) be a reference peptide (or protein) sequence database for sample $P$, ignoring protein abundance information[1], we

[1] We view protein database $DB$ or peptide database $PD$ as simplified representations (approximations) of a proteome sample before or after proteolysis respectively.

obtain,

$$\begin{aligned}
&\mathtt{P}(Y|MS2,\mathcal{T}) \propto \mathtt{P}(MS2|Y,\mathcal{T})\cdot \mathtt{P}(Y|\mathcal{T}) && (2.2)\\
&\qquad = \mathtt{P}(MS2|Y,\mathcal{T})\cdot \sum_{P}\mathtt{P}(Y|P,\mathcal{T})\cdot \mathtt{P}(P) && (2.3)\\
&\qquad \approx \mathtt{P}(MS2|Y,\mathcal{T})\cdot \sum_{PD}\mathtt{P}(Y|PD,\mathcal{T})\cdot \mathtt{P}(PD). && (2.4)
\end{aligned}$$

Conceptually, there are two pieces of prior information required to compute the posterior probability $\mathtt{P}(Y|MS2,\mathcal{T})$. First, the LC-MS/MS experimental protocol $\mathcal{T}$ needs to be modeled by a number of parameters, such as the mass accuracy of the mass spectrometer, the ionization methods, e.g. electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI), the fragmentation methods, e.g. collision induced dissociation (CID) or electron-transfer dissociation (ETD), and the proteolytic digestion protocol. The fragmentation method and the product ion mass accuracy are often considered by peptide search engines in their matching scoring scheme, and thus can be incorporated in the probability distribution $\mathtt{P}(MS2|Y,\mathcal{T})$ (or $\mathtt{P}(M|P,\mathcal{T})$ for protein inference). The precursor mass accuracy and the proteolysis protocol are currently considered by search engines to define the set of candidate peptides. In the Bayesian framework, along with the bias in the LC separation and ionization, they are modeled by the *peptide detectability* [34, 23] via the probability distribution $\mathtt{P}(Y|PD,\mathcal{T})$ or $\mathtt{P}(Y|P,\mathcal{T})$. The second type of prior information, encoded in the distribution $\mathtt{P}(P)$, reflects the prior information about the sample $P$, in particular the putative proteins and their abundances in the sample. For example, knowledge about the organism from which the sample is derived will enable the selection of the proper database of protein sequences to be used for peptide identification. Knowledge of the expected forms of post-translational modifications (PTMs) can be used to specify the fixed or variable PTMs when using peptide search engines. Finally, protein abundances (Figure 3.1a) that can be estimated from previous (transcriptomics [38] or proteomics) experiments on the same kinds of samples provide useful information to improve protein identification, although they are not explicitly considered in the current peptide/protein identification algorithms[2].

## 3. A BAYESIAN FRAMEWORK FOR PEPTIDE IDENTIFICATION

Based on the general Bayesian model described above, here we present a Bayesian interpretation for the common practices used in the peptide identification algorithms, which provides a computational framework to potentially improve peptide identification. It is worth pointing out that, although prior information is implicitly used in current peptide identification algorithms (both database search engines and *de novo* sequencing algorithms), the peptide identification problem has not been interpreted from a Bayesian perspective.

The Bayesian model for peptide identification is illustrated in Figure 3.1d. As discussed above, the information about either the experiment $\mathcal{T}$ or the sample $P$ (or database $PD$) can be used as priors. Below we will discuss the methods to incorporate several kinds of prior information into the model. First, we will show that protein database represents useful prior information in peptide identification. Then we will present different peptide prior distributions for peptide identifications given a deterministic protein database, and suggest that they can be utilized in both database searching and *de novo* sequencing algorithms to improve peptide identification. Next, for the cases where a complete protein database is unknown (e.g. for *de novo* sequencing), or uncertain (e.g. for the identification of peptides containing mutations or PTMs), we proposed several prior distributions over the space of peptide databases. Finally, after trying to combine the above two types of prior distributions (referred to as the *identification prior* and the *database prior*, respectively), we will propose the concept of *effective database size* as a useful measure for the interpretation of the impacts of prior knowledge on peptide identification.

### 3.1 Protein databases as prior information

There are at least three types of useful information embedded in a protein database. The first is the sequence information. When a protein sequence database $DB$ is used for peptide identification, we implicitly assume that a peptide from any protein in the database has some non-zero probability of being present in the sample, and the other peptides have zero probability and thus do not need be to considered for PSM scoring. The second is the abundance information. The database is biased against the low abundance proteins. This assumption is particularly important when the curated protein databases such as Swissprot [7] are used in peptide identification, because a protein to be included in these databases generally has a sufficiently high abundance (thus detected by some experimental methods) in some biological samples. Proteins at low expression levels may not be included in curated databases. The third piece of information is the sequence conservation information. The proteins in the databases tend to be more conserved across multiple species than the ones not in the database, because the existence of homologous proteins are often considered as a strong evidence for a protein to be real. Notably, the protein databases derived solely from *in silico* gene prediction software are rarely used in peptide identification as these databases could be much larger than curated databases while containing many mis-annotated proteins.

[2] The protein inference model MSBayesPro does explicitly estimate the abundance for candidate proteins, but informative prior distribution is not used to model protein abundances.

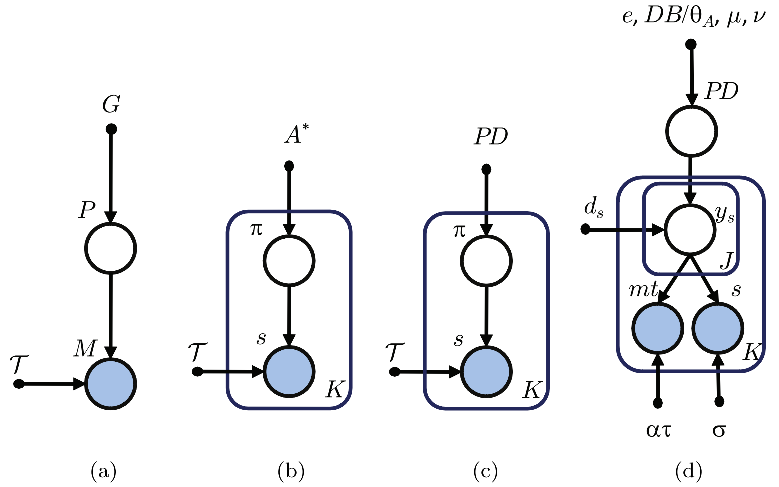

*Figure 3.1. Bayesian probabilistic models for peptide identifications from the LC-MS/MS analysis of proteome samples. (a) A high level graphical model for a* real *proteomics experiment; (b) A Bayesian interpretation of* de novo *peptide sequencing algorithms; (c) A Bayesian interpretation of database searching algorithms; (d) A model of Bayesian peptide identification. Graphical notations: dots, deterministic parameters; empty circle, random variables of interest; shaded circles, random variables with observed values. Symbolic notations: $P$, a proteome sample; $G$, prior information about the sample; $M$, the LC-MS/MS data; $\mathcal{T}$, LC-MS/MS platform; $A^*$, the set of all possible peptides formed from amino acids in $A$, which is the set of amino acids possibly with PTMs; $e$, parameters in the prior distribution of protein expression; $DB$, protein sequence database; $\theta_A$, amino acid frequencies, representing the frequency model for protein sequences; $\mu$: parameters in the prior distributions for amino acid mutations; $\nu$: parameters in the prior distributions for peptide modifications; $PD$: peptide sequence database; $y_s$ and $y$: random variables for a peptide to be fragmented in an MS/MS spectrum ($y_s$) or to be observed in a proteomics experiment ($y$); $d_s$ and $d$: per spectrum and per experiment detectabilities of peptide ions; $mt$, precursor mass and retention time associated with an MS/MS spectrum; $s$, an MS/MS spectrum; $\pi$: the random variable for a peptide in a sample; $J$ and $K$, the numbers of peptides and spectra considered in an analysis.*

Different peptide prior distributions can be used for peptide identification. In peptide search engines, peptides are deemed equally likely to be observed before peptide-spectra matchings are scored. Below, we present a Bayesian formulation of the peptide identification, in which non-uniform prior probabilities for different peptides may be employed. We note that in this section, we will follow the *peptide-centric* view of peptide identification (as shown in Figure 3.1d), which ignores the mutual dependence among peptides from the same proteins or different but co-occurrent proteins (e.g. proteins in the same biological pathway). More complex Bayesian models (as shown in Figure 4.1a–4.1c) that address these issues in the "real" proteomics experiment (as shown in Figure 3.1a) will be discussed in the next sections.

## 3.2 Prior distributions for peptide identification

In general, peptide identification is achieved by peptide search engines, in which each MS/MS spectrum $s_k$ is matched to each peptide in the peptide database $PD$.[3]

For simplicity, we assume individual MS/MS spectra are independent during the database searching. Thus, we have $\mathtt{P}(Y|MS2, P, \mathcal{T}) \propto \prod_k \mathtt{P}(s_k|y_k, PD, \mathcal{T}) \cdot \mathtt{P}(y_k|PD, \mathcal{T})$, where $y_k = (y_{k1}, \ldots, y_{kJ})$ is an indicator vector for peptide(s) that produces spectrum $s_k$, and $J$ is the number of peptides in the peptide database $PD$. Therefore, the goal of peptide identification becomes to estimate

$$(3.1)\qquad \mathtt{P}(y_k|s_k, PD, \mathcal{T}) = \frac{\mathtt{P}(s_k|y_k, PD, \mathcal{T}) \cdot \mathtt{P}(y_k|PD, \mathcal{T})}{\sum_{y_k \in \{0,1\}^J} \mathtt{P}(s_k|y_k, PD, \mathcal{T}) \cdot \mathtt{P}(y_k|PD, \mathcal{T})},$$

for each MS/MS spectrum $s_k$. The probability distribution $\mathtt{P}(y_k|PD, \mathcal{T})$ defined on $\{0,1\}^J$ is the prior distribution for the peptide identification.

Note that, in a typical LC-MS/MS proteomics experiment, an MS/MS spectrum can be produced from the fragmentation of a non-peptide molecule ion, a single peptide ion, a mixture of two or (rarely) more peptide ions (called a *mixture* spectrum), with $\sum_{j=1}^{J} y_{kj} \leq 2$ for most cases, implying that most of the indicator variables $y_{kj} = 0$. Furthermore, mixture spectra are known to comprise a small

[3] Here, we assume that a protein database $DB$ is converted to a set of peptides $PD$ to be used for the peptide identification from spectrum $s_k$ by using a computational procedure mimicking the protein digestion process. Precursor mass of an MS/MS spectrum can be applied as constraints to reduce the size of $PD$ (for details see Section 3.2.4).

portion (10%–20%) of all MS/MS spectra in a proteomics dataset [42, 6]. Therefore, in most proteomics analyses, only the PSM with the best score (i.e. top-ranked) for an MS/MS spectrum is retained after the database searching step. Albeit with small (posterior) probabilities, lower ranked PSMs can also be correct (whereas the top PSM is wrong). Hence, for the purpose of peptide identification, we argue that every candidate peptide should be assigned a probability. Although in practice lower ranked peptides are typically ignored for convenience in probabilistic PSM assessment algorithms such as PeptideProphet [19], in principle, the task of assigning probability to the top-ranked PSM of an MS/MS spectrum is best formulated as a subproblem of assigning probability for all PSMs for this MS/MS spectrum. Thus, in this paper, we will provide joint prior distribution for all candidate peptides (using the indicator variables $y_{kj}$ for all $j = 1, \dots, J$) for a spectrum $s_k$, rather than only for the top-ranked peptide[4].

To compute the posterior probability $\mathtt{P}(y_k|s_k, PD, \mathcal{T})$, we need to know both the likelihood $\mathtt{P}(s_k|y_k, PD, \mathcal{T})$ and the prior probability $\mathtt{P}(y_k|PD, \mathcal{T})$. We will focus on the modeling of the prior probabilities in this paper. Below we will present four specific types of distributions for $\mathtt{P}(y_k|PD, \mathcal{T})$, which fall into two categories: the *mutually exclusive* (ME) distributions, in which at most one peptide can be matched to an MS/MS spectrum, and the *independent* (ID) distributions, in which peptides are matched to an MS/MS spectrum independently (and thus more than one peptide is allowed to be matched to the same MS/MS spectrum). For each of the classes, we can either assign a uniform probability to different peptides in the database (and thus do not favor any specific peptide), or assign non-uniform probabilities to different peptides based on prior knowledge. As a result, we have four different prior distributions: the *identical mutually exclusive* (IME) prior distribution, the *detectability mutually exclusive* (DME) prior distribution, the *identical independent* (IID) prior distribution and the *detectability independent* (DID) prior distribution. Notably, all these prior distributions ignore the mutual dependence of peptides originated from a same protein, complying to the peptide-centric view of peptide identification (Figure 3.1d). These prior distributions, with properly selected parameters, can be used to estimate the posterior probabilities of peptides $\mathtt{P}(y_k|s_k, PD, \mathcal{T})$.

In experiments on complex biological samples, it is observed that a significant portion of the MS/MS spectra are not matched to any peptide in the database. Thus, for these four types of peptide identification prior distributions, we require a non-zero prior probability $p_-$ for a spectrum to be matched to none of the peptides in the database. The purpose is to allow an MS/MS spectrum to be produced from the fragmentation of an ion that does not correspond to a peptide in the database, but to a non-peptide molecule or a peptide absent from the database, e.g., a peptide containing unrecognized PTMs. Formally, we require the prior distributions to satisfy

$$\mathtt{P}(y_k = 0|PD, \mathcal{T}) = p_-. \tag{3.2}$$

Note that the parameter $p_-$ is database dependent. However, when the same database $PD$ is specified for the four types of prior distributions, we expect them to satisfy Equation 3.2 with same $p_-$. This will help us to better understand the relationship among the distributions.

### 3.2.1 Mutually exclusive prior distributions

Let $\mathcal{I}$ be the indicator function, i.e. $\mathcal{I}(true) = 1$ and $\mathcal{I}(false) = 0$. The IME prior distribution, denoted by $IME(p, J)$ with $p \in [0, 1]$ and $J \in \mathcal{N}$, is expressed as,

$$\mathtt{P}_{IME}(y_k|PD, \mathcal{T}) = (1-p) \cdot \mathcal{I}\left(\sum_{j=1}^{J} y_{kj} = 0\right) + \frac{p}{J} \cdot \mathcal{I}\left(\sum_{j=1}^{J} y_{kj} = 1\right). \tag{3.3}$$

The IME prior distribution is based on the assumption that an MS/MS spectrum has some probability $(1-p)$ of matching no peptide in the database, while it has equal prior probability $p/J$ to match to one and only one peptide in the database, where $J$ is the total number of peptides in the database. Applying the constraint of equation 3.2, we get $p = 1 - p_-$. Note that the assumption behind the IME prior distribution is implicitly used by most database searching engines, although it was not expressed as a prior distribution in a Bayesian model.

A DME prior distribution, denoted by $DME(l)$ with parameters $l = (l_1, \dots, l_J)$, satisfying $l_j \geq 0$ and $\frac{1}{J} \cdot \sum_j l_j \leq 1$, is expressed as,

$$\mathtt{P}_{DME}(y_k|PD, \mathcal{T}) = \left(1 - \frac{1}{J} \cdot \sum_j l_j\right) \cdot \mathcal{I}\left(\sum_j y_{kj} = 0\right) + \frac{1}{J} \cdot \prod_j l_i^{y_{ki}} \cdot \mathcal{I}\left(\sum_j y_{kj} = 1\right). \tag{3.4}$$

As in the IME prior distribution, PSMs in DME prior are assumed to be mutually exclusive from each other. However, to model the identification bias, peptide-specific parameters $l = (l_1, \dots, l_J)$ are introduced, each for a specific PSM between spectrum $s_k$ and peptide $\pi_j$, indicated by $y_{kj} = 1$ (and $y_{kj'} = 0$ for $j' \neq j$). After introducing the IID and DID prior distributions below, we will give an alternative parameterization of DME distributions in which the peptide detectabilities are used as parameters (see Section 3.2.3). We

[4] Note that when we say the peptide(s) are identified for a spectrum, we really mean the peptide ion(s) of specific charges are identified. However, to keep the notation simple, throughout this paper, we use $j = 1, \dots, J$ to indicate the peptides, ignoring their charges.

will show that the parameters $l_j$ are closely related to the peptide detectabilities [34].

### 3.2.2 Independent prior distributions

By definition, when an IME or DME prior distribution is used, the probability for two or more peptides to be identified for a single spectrum is 0. To eliminate this limitation, we define the IID prior distribution, denoted by $IID(p_0, J)$, as

$$\mathtt{P}_{IID}(y_k|PD, \mathcal{T}) = \prod_j (1-p_0)^{1-y_{kj}} \cdot {p_0}^{y_{kj}}. \tag{3.5}$$

It allows an MS/MS spectrum to be matched to any number (from 0 to $J$) of peptides in the database, and thus allows multiple peptides to be identified for the *mixture spectra*, i.e. those spectra resulted from the fragmentation of the mixture of multiple peptide ions. The IID prior distribution assumes that each peptide in the database has an equal and small probability $p_0$ to be matched to the spectrum $s_k$. Applying the constraint of equation 3.2, we get $p_0 = 1 - p_-^{1/J}$. Remember that for the IME distribution we have $p = 1 - p_-$. If we set $p_0 = 1 - (1-p)^{1/J} \approx -\frac{1}{J} \cdot \ln(1-p)$, the distributions $IID(p_0, J)$ and $IME(p, J)$ are approximately equivalent. As we mentioned above, IID prior allows mixture spectra while IME prior does not. However, we can show that when $p \to 0$ (i.e. $p_- \to 1$), $\mathtt{P}_{IID}(\sum_j y_{kj} > 1|PD, \mathcal{T}) = 1 - p_-[1 + J \cdot p_0/(1-p_0)] \approx 1 - (1-p) \cdot [1 - \ln(1-p)] = O(p^2)$ can be ignored, i.e. when the fraction of the identifiable spectra is small, the fraction of mixture spectra is negligible for the IID prior.

The DID prior distribution, denoted by $DID(d)$ with parameters $d = (d_1, \ldots, d_J)$ and $d_j \in [0, 1]$, can be expressed as

$$\mathtt{P}_{DID}(y_k|PD, \mathcal{T}) = \prod_j (1-d_j)^{1-y_{kj}} \cdot d_j^{y_{kj}}. \tag{3.6}$$

Similar to the DME distribution, the DID prior distribution is an extension of the IID by incorporating the peptide specific information for each peptide in the prior distribution. The DME prior distribution considers peptide detectability $d_j$ [34, 23] for each peptide $j$. It should be pointed out that, 1) to be consistent with the primary goal of the peptide identification (Figure 3.1d), we consider here the peptide detectability $d_j$ as the probability of a peptide to be matched to a single MS/MS spectrum $s_k$ rather than to any of a set of MS/MS spectra in a dataset[5], and thus it is a much smaller value; and 2) the peptide detectability should be predicted only from the sequence of the peptide but not from the sequence and quantity of the protein from which the peptide is generated, because we generally do not have such information in the peptide identification step. This is different from the protein inference task, in which peptide detectability is predicted using the information of the protein sequence as well as the protein abundance (Figure 4.1b).

### 3.2.3 Relationships between the ME and ID prior distributions

We can see that the DID prior distribution is a generalization of the IID prior distribution by allowing different peptides to have different prior probabilities of being identified by MS/MS spectra. If we assume $d_j = p_0$ for all $j = 1, \ldots, J$ in the DID distribution, then it is reduced to IID. By solving $\mathtt{P}_{DID}(y_k = 0|PD, \mathcal{T}) = p_-$, we get $\prod_{j=1}^J (1-d_j) = p_-$, which can be re-written as $\sum_{j=1}^J \ln(1-d_j)/J = \ln(1-p_0)$, and approximately $p_0 = \sum_{j=1}^J d_j/J$ when $p_0 \to 0$ and $d_j \to 0$. This suggests that the IID prior distribution can be viewed as an approximation of DID. Similarly, the IME distribution is a special case of the DME distribution. To better understand the relationship between DID and DME, we first define the following concepts.

**Definition 3.1** (Degree of a distribution)**.** For a positive integer $J$ and an arbitrary distribution $\Psi$ defined on $\{0,1\}^J$ – the set of indicator vectors, a number $n$ is called the *degree* of the distribution, denoted as $deg(\Psi) = n$, iff $\mathtt{P}_\Psi(\sum_{j=1}^J y_j = n) > 0$ and $\mathtt{P}_\Psi(\sum_{j=1}^J y_j > n\ |\theta) = 0$.

The terminology is inspired by the *degree of a polynomial* concept and an equivalent definition of $DID(d)$ as a polynomial, i.e. $DID(y_k; d) = \prod_j [1 - d_j + (2 \cdot d_j - 1) \cdot y_{kj}]$. It can be shown that each distribution $\Psi$ defined on $\{0,1\}^J$ has a corresponding polynomial formulation. However, in general, the degree of the corresponding polynomial does not equal the degree of the distribution.

Intuitively, $deg(\Psi)$ is the maximum number of the peptides allowed for a mixture spectrum if $\Psi$ is the prior distribution for peptide identification from this spectrum. One critical difference between the DID and the PID distributions is that $\mathtt{P}_{DID}(\sum_j y_{kj} > 1) > 0$ while $\mathtt{P}_{DME}(\sum_j y_{kj} > 1) = 0$. Formally, we see that $deg(DID)$ equals the number of peptides in the database $PD$ with non-zero detectabilities, while $deg(DME)$ equals 1. Meanwhile, as we discussed earlier, the IID and IME prior distributions are approximately equivalent when $p \to 0$, because $\mathtt{P}_{IID}(\sum_j y_{kj} > 1|PD, \mathcal{T}) = O(p^2)$. Similarly, we can prove that $\mathtt{P}_{DID}(\sum_j y_{kj} > 1|PD, \mathcal{T}) = O((1-p_-)^2)$ when $\mathtt{P}_{DID}(y_k = 0|PD, \mathcal{T}) = p_- \to 1$, and thus, the DID and DME distributions also approximate each other when $p_- \to 1$.

Based on this observation, we define the 1st-order approximation of the DID distribution, i.e. to only keep the 0th- and 1st-degree terms of the expanded DID distribution as defined in equation 3.6. This will result in a mutually exclusive distribution with peptide detectabilities $d = (d_1, \ldots, d_J)$ as

[5] To be strict, the notation $d_s$ with subscript $s$ is used to denote the per-spectrum peptide detectability as in Figure 3.1d. In the text, we ignore the subscript $s$ for simplicity.

parameters. We denote this distribution as $DME(d)$.

$$\begin{aligned}(3.7)\quad & \mathtt{P}_{DME}(y_k|PD,\mathcal{T}) \\ &= \left[1+\sum_j (d_j/(1-d_j))\right]^{-1}\cdot\left[\mathcal{I}\left(\sum_j y_{kj}=0\right)\right. \\ &\left.+\mathcal{I}\left(\sum_j y_{kj}=1\right)\cdot\prod_j (d_j/(1-d_j))^{y_{kj}}\right].\end{aligned}$$

Comparing with the previous definition of the DME prior distribution in equation 3.4, we can find the transformation between the alternative parameterizations of the DME distribution as,

$$(3.8)\qquad d_j = \frac{l_j}{l_j+\sum_{m=1}^J (1-l_m)},$$

$$(3.9)\qquad l_j = \frac{d_j/(1-d_j)\cdot J}{1+\sum_{m=1}^J (d_m/(1-d_m))}.$$

With this transformation, the two parameterizations of DME distribution are equivalent if a fixed database PD is used. Unfortunately, the transformation cannot be applied simultaneously to different databases unless DME is reduced to IME. Accordingly, if we consider different $PD$s in the peptide identification (e.g. as in Section 3.4), the constraint $\mathtt{P}(y_k = 0|PD,\mathcal{T}) = p_-$ does not hold for DME or DID distributions unless they are reduced to IME or IID. Observe that the transformation between the parameters $d_j$ and $l_j$ for DME distribution are database dependent (i.e. if we assume peptide detectability $d_{s\pi}$ for peptide $\pi$ to be independent of the database where $\pi$ is in, then $l_{s\pi}$ will take different values when $\pi$ is in different $PD$; and *vice versa*), unless $\sum_{t=1}^J (1-l_t)$ or $\sum_{t=1}^J (d_t/(1-d_t))$ is constant for different databases, which turns out to be equivalent to requiring $\mathtt{P}(y_k = 0|PD,\mathcal{T})$ to be constant for all databases. To address the issue, we set $\mathtt{P}(y_k = 0|PD,\mathcal{T}) = p_-$ for one database $PD_0$ that is used to build the predictors for $d_j$ and $l_j$, and then we can rewrite the transformation by applying $\mathtt{P}(y_k = 0|PD,\mathcal{T}) = p_-$ for $PD_0$ and obtain a new transformation below, which generalizes the parameters to the peptides that are not in $PD_0$ and thus is database independent.

$$(3.10)\qquad d_j = \frac{l_j}{l_j + J\cdot p_-},$$

$$(3.11)\qquad l_j = \frac{d_j\cdot J\cdot p_-}{1-d_j}.$$

We note that, with this transformation, the $l$ and $d$ parameterizations of DME distribution are no longer equivalent (expect for the database $PD_0$) but only approximate each other. Below, we will use these two parameterizations for DME alternatively, depending on which one is more convenient.

### 3.2.4 Using retention time and precursor mass as prior information

The retention time at which a fragmentation spectrum is acquired and the precursor mass can also be modeled through the peptide identification prior distribution, to further specify the putative peptides to be matched to the spectrum[6]. Formally, to combine the precursor mass $m_k$ and the retention time $t_k$ together with the fragmentation $s_k$ so as to achieve better posterior probability for peptide identification, we apply the Bayes rule under the assumption of *conditional independence* among $s_k$, $m_k$ and $t_k$, given $y_k$, $PD$, and experiment $\mathcal{T}$.

$$\begin{aligned}(3.12)\quad & \\ & \mathtt{P}(y_k|s_k,m_k,t_k,PD,\mathcal{T}) \\ &= \frac{\mathtt{P}(s_k|y_k,PD,\mathcal{T})\cdot\mathtt{P}(y_k|m_k,t_k,PD,\mathcal{T})}{\sum_{y_k\in\{0,1\}^J}\mathtt{P}(s_k|y_k,PD,\mathcal{T})\cdot\mathtt{P}(y_k|m_k,t_k,PD,\mathcal{T})}.\end{aligned}$$

We see that $\mathtt{P}(y_k|m_k,t_k,PD,\mathcal{T})$ may replace $\mathtt{P}(y_k|PD,\mathcal{T})$ to be the prior distribution for peptide identification. It can be shown that, under reasonable assumption for $\mathtt{P}(m_k|y_k,PD,\mathcal{T})$ and $\mathtt{P}(t_k|y_k,PD,\mathcal{T})$, if $\mathtt{P}(y_k|PD,\mathcal{T})$ is a DME or DID distribution, then $\mathtt{P}(y_k|m_k,t_k,PD,\mathcal{T})$ is still a DME or DID distribution, only with the parameters $d_j = \mathtt{P}(y_{kj} = 1|PD,\mathcal{T})$ replaced by $d_{kj} = \mathtt{P}(y_{kj} = 1|m_k,t_k,PD,\mathcal{T})$, for $j = 1,\dots,J$ and any $k$.

**Proposition 3.2** (mtDME and mtDID distributions)**.** *If a DME or DID prior distribution is used in peptide identification with per spectrum peptide detectabilities $(d_j)_{j=1,\dots,J}$ as parameters, i.e. $\mathtt{P}(y_k|DB,\mathcal{T}) = DME(y_k;d)$ or $\mathtt{P}(y_k|DB,\mathcal{T}) = DID(y_k;d)$, further assuming the likelihood function for precursor mass can be expressed as $\mathtt{P}(m_k|y_k,PD,\mathcal{T}) = \prod_{j:y_{kj}=1} f_+(m_k;\alpha_j)\cdot\prod_{j:y_{kj}=0} f_-(m_k)$ (where $f_+$ and $f_-$ represent the likelihood of correctly and incorrectly matched peptides to have the precursor mass $m_k$, respectively), and similarly, the likelihood function for retention time can be expressed as $\mathtt{P}(t_k|y_k,PD,\mathcal{T}) = \prod_{j:y_{kj}=1} g_+(t_k;\tau_j)\cdot\prod_{j:y_{kj}=0} g_-(t_k)$, then the peptide identification still follow a DME or DID prior distribution (denoted by mtDME and mtDID, respectively), with the updated detectabilities $(d_{kj})_{j=1,\dots,J}$ as parameters, where*

$$d_{kj} = \frac{f_+(m_k;\alpha_j)\cdot g_+(t_k;\tau_j)\cdot d_j}{f_-(m_k)\cdot g_-(m_k)\cdot(1-d_j)+f_+(m_k;\alpha_j)\cdot g_+(m_k;\tau_j)\cdot d_j},$$

*for $j = 1,\dots,J$.*

*Proof.* Below we prove the proposition for the case of DME distribution. A similar argument can also be applied to the case of DID distribution. Let $y_{kj} = 1$ be an indicator vector with zeros for all but one variable (peptide) indexed by $j$,

[6] Note that incorporating retention time and precursor mass information into the peptide identification scoring has been discussed in [31] among others, although not within a Bayesian framework.

and let $y_k = 0$ be an indicator vector with all zero values. Because $m_k$ and $t_k$ are independently observed from $s_k$ given $y_k$, $PD$ and $\mathcal{T}$, we have,

$$\begin{aligned}
&\mathtt{P}_{mtDME}(y_k|m_k, t_k, PD, \mathcal{T})\\
&\quad\propto \mathtt{P}_{DME}(y_k|PD, \mathcal{T}) \cdot \mathtt{P}(m_k|y_k, PD, \mathcal{T}) \cdot \mathtt{P}(t_k|y_k, PD, \mathcal{T})\\
&\quad\propto \left[\mathcal{I}\left(\sum_{j=1}^{J} y_{kj} = 0\right) + \mathcal{I}\left(\sum_{j=1}^{J} y_{kj} = 1\right)\right.\\
&\qquad\left.\cdot \prod_{j=1}^{J} (d_j/(1-d_j))^{y_{kj}}\right] \cdot \mathtt{P}(m_k|y_k, PD, \mathcal{T})\\
&\qquad\cdot \mathtt{P}(t_k|y_k, PD, \mathcal{T})\\
&\quad\propto \mathcal{I}\left(\sum_{j=1}^{J} y_{kj} = 0\right) + \mathcal{I}\left(\sum_{j=1}^{J} y_{kj} = 1\right)\\
&\qquad\cdot \prod_{j=1}^{J} \left(\frac{d_j \cdot f_+(m_k;\alpha_j) \cdot g_+(t_k;\tau_j)}{(1-d_j) \cdot f_-(m_k) \cdot g_-(t_k)}\right)^{y_{kj}}.
\end{aligned}$$

Comparing with $DME(y_k; d_k)$, we find that $\frac{d_{kj}}{1-d_{kj}} = \frac{d_j \cdot f_+(m_k;\alpha_j) \cdot g_+(t_k;\tau_j)}{(1-d_j) \cdot f_-(m_k) \cdot g_-(t_k)}$, which gives updated detectability as

(3.13)

$$d_{kj} = \frac{f_+(m_k;\alpha_j) \cdot g_+(t_k;\tau_j) \cdot d_j}{f_-(m_k) \cdot g_-(m_k) \cdot (1-d_j) + f_+(m_k;\alpha_j) \cdot g_+(m_k;\alpha_j) \cdot d_j}.$$

Therefore, we have $\mathtt{P}_{mtDME}(y_k|m_k, t_k, PD, \mathcal{T}) = DME(y_k; d_{kj})$. □

According to the above proposition, we can use the experimentally observed precursor mass and retention time to refine peptide detectability $d_j$, and thus give an updated prior distribution for peptide identification in the same dataset. We note that while $d_j$ is the same for all MS/MS spectra, $d_{kj}$ is specific for each spectrum $s_k$.

### 3.2.5 How does a prior distribution impact peptide identification?

We pointed out earlier that the IME prior distribution is implicitly used by most database search engines, because 1) they usually do not consider mixture spectra; and 2) applying the IME prior will not change the PSM ranks for the same MS/MS spectrum. As a result, the performance of a scoring scheme in the peptide identification will not be affected by applying IME prior distribution.

On the contrary, both DID and DME distributions assign peptide-specific prior probabilities for peptide identification. Although these prior distributions have not been used in practice, we believe they will provide a simple yet realistic model for improving the peptide identification from fragmentation spectra (including mixture spectra), assuming peptides are independently eluted and the ions are *randomly* selected for fragmentation. We stress that DID and DME prior distributions may improve peptide identification in two ways: 1) it may re-rank the top-ranked PSMs, each from a specific MS/MS spectrum; and 2) it may also re-rank the PSMs for a single spectrum. As a result, a peptide (or peptides for a mixture spectrum) with lower rank of likelihood may become top ranked by the posterior probabilities.

## 3.3 Peptide identification priors for de novo peptide sequencing

At first glance, a *de novo* peptide sequencing algorithm does not seem to rely on any protein database as prior information. However, they indeed use it implicitly. Instead of using a peptide database $PD \subseteq A^*$, it uses directly $A^*$, which is the set of all possible peptides formed from amino acids in $A$, the alphabet of amino acids including the ones containing possible PTMs. The resulting peptide database $PD_A$ for a spectrum $s_k$ is typically much larger than that used in the database searching algorithm. In practice, the *de novo* sequencing usually uses a sophisticated algorithm (e.g. dynamic programming) that does not exhaustively examine the peptides in database $A^*$, and thus is much more efficient. However, consider a brute force algorithm that matches every peptide in $A^*$ against every spectrum $s_k$, the matching results will be the same as the *de novo* sequencing algorithm using the same scoring. Hence, if we consider $A^*$ as the database[7] for *de novo* peptide sequencing, then *de novo* sequencing and database searching approaches are not different in principle. We can thus analyze *de novo* sequencing algorithms using exactly the same Bayesian framework and the same peptide identification prior distributions outlined above as used for database searching algorithms.

The relationship among *de novo* sequencing, database searching, and Bayesian peptide identification is evident from the perspective of the probabilistic graphical model representations, given in Figure 3.1b–3.1d. For both *de novo* sequencing and database searching, a single peptide is sampled from a given peptide database with equal probability and then matched to the MS/MS spectrum, whereas for the Bayesian peptide identification model in Figure 3.1d, a non-deterministic peptide database $PD$ is allowed (see Section 3.4), and multiple peptides can be sampled together at non-equal probabilities for one MS/MS spectrum. The *de novo* peptide sequencing (Figure 3.1b) is a special case of the database searching (Figure 3.1c), when $PD = A^*$; and the database searching can be viewed as a special case of Bayesian peptide identification, when $\mathtt{P}(\mathbf{PD}) = \mathcal{I}(\mathbf{PD} = PD)$ and a $IME$ prior distribution is used. Therefore, the Bayesian peptide identification model illustrated in Figure 3.1d provides a unified interpretation for different peptide identification approaches.

[7] Note that we will provide an alternative and better treatment later by modeling the peptide database as a random variable for *de novo* sequencing.

### 3.3.1 Impact of large databases

The prominent difference between *de novo* sequencing and database searching approaches is that the database size $J$ is significantly larger (theoretically infinite) for *de novo* peptide sequencing algorithms. The number of peptides increases dramatically with the increasing size of alphabet $A$ of amino acids (and the ones with PTMs) on the order of $O(||A||^L)$ with $L > 7$ typically being the peptide length. Large $J$ will lead to low *identifiability* of the true peptide, where the *identifiability* is referred to as the power of a PSM scoring scheme to identify a specific true peptide with high confidence from a set of MS/MS spectra. Because the prior probability for a specific peptide $j$ to be identified, i.e. $\mathtt{P}(y_{kj} = 1, y_{k\bar{j}} = 0|PD, \mathcal{T})$, approaches 0 for both the PME prior distribution (equation 3.3), with $\mathtt{P}(y_{ki} = 1, y_{k\bar{j}} = 0|PD, \mathcal{T}) = \frac{1-p_-}{J}$, and the IID prior distribution (equation 3.5), with $\mathtt{P}(y_{ki} = 1, y_{k\bar{j}} = 0|PD, \mathcal{T}) \approx -\frac{1}{J} \cdot \ln p_-$. It can be shown that if new peptides are randomly added to the database, and a fixed fraction of the newly added peptides match spectrum $s_k$ with non-zero likelihood $\mathtt{P}(s_k|y_k, PD, \mathcal{T})$ for $y_k \neq 0$, and then $\lim_{|PD|\to\infty} \mathtt{P}(y_{kj} = 1, y_{k\bar{j}} = 0|s_k, PD, \mathcal{T}) = 0$. This indicates that if we assume a uniform prior probability distribution over all putative peptides, the *de novo* sequencing algorithms may have very small power to identify the true peptides, due to the very large number of the putative peptides. It is worth pointing out that, since the effective database size $\mathcal{N} < \infty$ after filtering for precursor mass and possibly retention time for a specific MS/MS spectrum, the sensitivity of the *de novo* sequencing is the result of a battle between the large database size (hence low information content in the prior) and the useful information content in a MS/MS spectrum. If a spectrum is informative enough to tell apart the true peptide(s) from very similar false peptides, the *de novo* sequencing using IID or IME priors can still work. However, unlike for DNA sequencing techniques, the signal/noise ratio in MS/MS spectra are often not high enough to accomplish this goal.

Interestingly, these observations regarding database size does not apply to the DID and DME prior distributions, which consider the peptide-specific prior probabilities. This is obviously true, because a properly designed DID (DME) prior distribution for a *de novo* sequencing algorithm can effectively approximate either IME or IID prior distributions used in the database searching methods which use a much smaller peptide database $PD$: we can simply set $d_j = 0$ for peptides not in $PD$. Of course, this is not a practical approach for assigning prior in real applications since $PD$ is typically unknown (and assumed to be $A^*$) when *de novo* algorithms are used. However, the peptide detectability prediction and other information (such as precursor mass and retention time information as discussed above, and the amino acid frequencies in real proteins and the prior probabilities of PTM occurrences that will be discussed below) may be combined to define better prior distributions, and thus the power of the *de novo* peptide sequencing can be improved. In summary, it is possible to overcome the low identifiability limitation of the *de novo* sequencing algorithms by applying informative DID/DME prior distributions over the peptides.

## 3.4 Prior distributions over databases

For the prior distributions discussed so far, we assumed the database is pre-determined: the IME, DME, IID and DID prior distributions are defined on a given peptide database $PD$, and the *de novo* peptide sequencing uses $PD = A^*$. Although in the DME and DID prior distributions, the peptide-specific bias of peptide identification in an LC-MS/MS experiment is modeled by the peptide detectability, all peptides in the database are assumed to be equally likely present in the sample. Besides, the abundances of these peptides are not modeled.

In practice, however, the *correct* peptide database can be unknown, as for the *de novo* peptide sequencing, or *uncertain*, as for the identification of mutations and PTMs in proteins when the mutation forms and PTMs of the proteins are not known for sure to be present. When the protein sequences are predicted from genomic or transcriptomic sequences, some sequences in the database are more likely to be real while some others are more likely to be erroneous due to prediction mistakes. Such information may be captured by modeling $\mathtt{P}(PD)$, the prior distribution of the peptide database $PD$.

An ideal protein database for protein identification from a biological sample should have a specific probability distribution for the abundances of all putative proteins, some of which may have high probability to be abundant (e.g., the ones in the curated database) whereas the others may have low probability to be abundant (e.g., the ones predicted from the genomic sequences). In practice, this is hardly achievable, although it can be approached by incorporating previous transcriptomics/proteomics studies of the same sample.

### 3.4.1 Amino acid frequencies based prior distribution

For the *de novo* peptide sequencing, more prior information can be employed other than peptide detectability. A simple observation is that, in all organisms, the amino acid occurrences in the *bona fide* proteins follow a non-uniform distribution: some amino acids occurred more often than the others. As a result, a peptide $\pi \in A^*$ with multiple rare amino acids is unlikely to be real. For a tryptic peptide $\pi = \pi_{(1)} \cdot \pi_{(2)} \cdot \ldots \cdot \pi_{(L)}$ of length $L$, we define,

(3.14)

$$\mathtt{P}(\pi|\theta_A) = \mathcal{I}(\pi_{(L)} \in \{K, R\}) \cdot \theta_{\pi_{(L)}} \cdot \prod_{j=1}^{L-1} [\mathcal{I}(\pi_{(j)} \notin \{K, R\}) \cdot \theta_{\pi_{(j)}}].$$

The probability distribution is defined in the space of all peptides $A^*$, where $\theta_A = (\theta_a)_{a \in A}$ represent the expected probabilities for amino acids in $A$ to occur in real proteins. We can see that the distribution 3.14 is an extension of the geometric distribution defined on $\{0,1\}^*$. It assigns probability 0 to any mis-cleavage or non-tryptic peptide. The distribution can be further generalized to model mis-cleaved or non-tryptic peptides. To do that, we use a Markov chain model of peptides with parameters $\theta_A$ and $\tau_A = (\tau_a)_{a\in A}$, where $\tau_a$ represents the transition probability from an amino acid $a \in A$ to the C-terminal, while $\theta_a$ is the transition probability from any amino acid to $a$. We then can obtain a prior distribution for peptides with any C-terminal amino acid specificity (resulted from the digestion of any protease or other peptide truncation mechanisms [1]) as,

$$\mathtt{P}(\pi|\theta_A,\tau_A) = \tau_{\pi(L)} \cdot \theta_{\pi_{(L)}} \cdot \prod_{j=1}^{L-1} [(1-\tau_{\pi(j)}) \cdot \theta_{\pi_{(j)}}]. \tag{3.15}$$

The distribution in equation 3.14 is a special case of the one in equation 3.15 with $\tau_{a\in\{K,R\}} = 1$ and $\tau_{a\notin\{K,R\}} = 0$. In general, $\tau_A$ signifies the protease specificity, while $\theta_A$ signifies the amino acid frequencies in real peptides[8]. Compared to 3.14, 3.15 allows all peptides in $A^*$ to be present when $\forall_{a\in A}\tau_a \neq 0$, although shorter peptides are always preferred.

Given a well-defined prior distribution for peptides, we can then define the *amino acid frequency* (AF) prior distribution as

$$\mathtt{P}_{AF}(PD|\theta_A,\tau_A) = \prod_{j=1}^{J} \mathtt{P}(\pi_j|\theta_A,\tau_A), \tag{3.16}$$

for a peptide database $PD = (\pi_j)_{j=1,\ldots,J} \in (A^*)^J$ of size $J$. Again, here we neglect the mutual dependency among peptides from the same protein. Of course, protein sequences are not available for *de novo* sequencing.

### 3.4.2 Prior distributions for peptides containing mutations and PTMs

The problems of identifying the mutations and PTMs in peptides remain challenging in proteomics. These problems can be approached by either database searching or *de novo* peptide sequencing algorithms. *De novo* sequencing of peptides containing mutations is not different from the conventional *de novo* peptide sequencing, because the search space remains the same (as the set of all peptides). The *de novo* identification of PTMs in peptides can be viewed as the same problem as the *de novo* peptide sequencing on an expanded amino acid alphabet $A'$, which includes the modified amino acids. For database searching algorithms, to identify the mutated and modified peptides, an expanded peptide database should be constructed, explicitly or implicitly, containing the mutated or modified peptides in addition to the original peptides. Database searching algorithms can then be applied in the same way as used for the identification of unmodified peptides [3, 4]. An issue of this non-Bayesian approach is that the mutated and modified peptides are deemed equally likely to be present *a priori* as the unmodified peptides, and different mutations and PTMs in the same peptides are also deemed equally likely. In practice, however, the faction of fragmentation spectra resulted from peptides containing mutations or PTMs is either much smaller (e.g., in a shotgun proteomics experiment) or much greater (e.g., when the sample is enriched for certain PTM such as phosphorylation, or when the protein database from a different species is used for the searching) than the fraction of spectra from unmodified peptides. In some cases, only a few MS/MS spectra correspond to form modified peptides, while the number of putative modified peptides in the expanded database can be several orders of magnitude larger than the unmodified peptides[9].

This issue may be addressed within a Bayesian framework, in which a prior distribution of the modified and unmodified peptides can be obtained from previous experimental analysis of related samples[10]. Ideally, the prior distribution should incorporate two types of prior information: 1) the propensity of the mutations and PTMs, i.e. how likely the mutations and PTMs occur in each site of a specific peptide; and 2) the dependency between the mutant/PTM forms and the unmodified peptides, because the modified and unmodified peptides tend to occur mutually exclusively for a given site of a protein in real samples.

For an unmodified peptide $\pi_j \in PD_0$ (where $PD_0$ is the reference database of unmodified peptides), we denote the set of peptides mutated from it as $\Delta_\mu\pi_j \subseteq A^*$, and the set of peptides modified from it as $\Delta_\nu\pi_j \subseteq A'^*$, where $A'$ is the expanded alphabet including the modified amino acids, $\mu$ and $\nu$ are the parameters for modeling the processes of amino acid mutations and post-translational modifications[11], respectively. The propensity of amino acid mutations can then be modeled by the probability distribution $\mathtt{P}(\pi'|\pi,\mu)$ for all $\pi \in A^*$ and $\pi' \in \Delta_\mu\pi$. And similarly, the propensity of PTMs can be modeled by $\mathtt{P}(\pi'|\pi,\nu)$ for $\pi' \in \Delta_\nu\pi$. For brevity, we will not discuss the specific forms of $\mathtt{P}(\pi'|\pi,\mu)$ and $\mathtt{P}(\pi'|\pi,\nu)$ in this paper. But we comment that they are conceptually easy to derive based on the amino acid substitution rate matrix (for $\mathtt{P}(\pi'|\pi,\mu)$) [17] and some PTM predictors (for $\mathtt{P}(\pi'|\pi,\nu)$) [18], respectively. Once they are determined, these probability distributions can be incorporated as priors into database searching or *de novo* sequencing algorithms for the identification of mutations and PTMs.

[8] In order to model the C-terminal peptides in proteins, we need to modify $\mathtt{P}(\pi|\theta_A,\tau_A)$ so that $\tau_{\pi(L)} = 1$ for all C-terminal amino acid $\pi(L)$. A mixture distribution can be used to model both protein C-terminal peptides and other peptides.

[9] Due to the same reason, the identification of mutations and PTMs is also computationally expensive.

[10] Note that fitting the prior distribution from the current dataset is feasible only as a post-processing step within the Bayesian framework.

[11] Note that $\pi_j \in \Delta_\mu\pi_j$ and $\pi_j \in \Delta_\nu\pi_j$.

For the identification of mutations in an unmodified peptide $\pi_j \in PD_0 = (\pi_j)_{j=1,\ldots,J}$, we assert that there is one and only one peptide[12] in $\Delta_\mu \pi_j$ that can be present in the *target* database $PD = (\pi'_j)_{j=1,\ldots,J}$, i.e. either the original peptide or a mutant of the peptide. As a result, we can define the mutDB prior distribution over the target peptide database as,

$$\mathtt{P}_{mutDB}(PD|PD_0, \mu) = \prod_{i=1,\ldots,J} \mathtt{P}(\pi'_i|\pi_i, \mu). \tag{3.17}$$

Note that, unlike the AF prior distribution used for *de novo* peptide sequencing, mutDB prior distribution is dependent on a reference peptide database $PD_0$. We see that the sample space of $PD$ is $\Delta_\mu \pi_1 \times \cdots \times \Delta_\mu \pi_J$, which is a set of databases of the same size as the reference database $PD_0$. Although the size of the database $PD$ is $J$, the size of $\cup_{\pi \in DB} \Delta_\mu \pi$, the set of peptides with non-zero marginal probabilities, is much larger.

In the PTM identification, more than one PTM forms of the unmodified peptide $\pi_i$ can be present in $\Delta_\nu \pi_i$, including the unmodified peptide itself. However, the total abundance of all PTM forms should be summed up to a fix value, assuming all peptides are at the same abundances[13], the *standard* abundance, for the peptide identification. Therefore, distinct from all prior distributions discussed so far, the prior distribution for PTM peptide database models not only peptide sequences but also their abundances. Assuming that the standard abundance of a peptide corresponds to $n_0$ copies of peptides in the sample, and assuming a multinomial distribution for the copy numbers of all possible PTM forms according to the PTM propensity defined by $\mathtt{P}(\pi'|\pi, \nu)$, we obtain the modDB prior distribution over all modified peptide databases as

$$\begin{aligned} &\mathtt{P}_{modDB}(q = (q_{jh})_{jh}|PD_0, \nu, n_0) \\ &\quad = \prod_{j=1,\ldots,J} n_0! \cdot \mathcal{I}(\sum_h q_{jh} = n_0) \cdot \prod_{h:\pi_{jh} \in \Delta \pi_j} \frac{\mathtt{P}(\pi_{jh}|\pi_j, \nu)^{q_{jh}}}{q_{jh}!}, \end{aligned} \tag{3.18}$$

where $q_{jh}$ represents the abundance of the $h$-th PTM form $(\pi_{jh})$ of the peptide $\pi_j \in PD_0$. Note that the sample space of $q = (q_{jh})$ is a subset of $\mathcal{N}^{\sum_j |\Delta_\nu \pi_j|}$. Although the two prior distributions, modDB (which models peptide abundances) and mutDB (which models peptide presence only), are different in general, modDB is reduced to mutDB when $n_0 = 1$.

## 3.5 Combining database priors with peptide identification priors

We argue that the best practice for Bayesian peptide identification should combine all prior information for peptide identification, including 1) the bias of amino acid frequency in real peptides; 2) the mutation and PTM propensity; 3) the sequence database of unmodified peptides and preferably also the probability that each of peptide/protein is real; 4) the retention time and mass of the precursor ion of each MS/MS spectrum; and 5) peptide detectabilities. The first three can be modeled in the protein sample/database prior distribution, whereas the other two can be modeled in the peptide identification prior distribution.

These two categories of prior distributions model completely different aspects of a proteomics experiment. Database prior distributions model the presence of peptides in the sample, whereas the peptide identification prior distributions model the possibility of peptides to be captured in a proteomics experiment. Given a peptide database prior distribution $\mathtt{P}(PD|\theta)$ parameterized by $\theta$, and a peptide identification prior distribution $\mathtt{P}(y_k|PD, \mathcal{T})$ or $\mathtt{P}(y_k|m_k, t_k, PD, \mathcal{T})$ for an MS/MS spectrum, a natural way of combining the two is to marginalize out $PD$ and obtain $\mathtt{P}(y_k|\theta, \mathcal{T}) \triangleq \sum_{PD \subseteq A^*} \mathtt{P}(y_k|PD, \mathcal{T}) \cdot \mathtt{P}(PD|\theta)$. For example, for the *de novo* peptide sequencing, combining the AF prior distribution and the DID prior distribution, we can define AF-DID prior distribution as $\mathtt{P}_{AF\text{-}DID}(y_k|\theta_A, \mathcal{T}) = \sum_{PD \subseteq A^*} \mathtt{P}_{DID}(y_k|PD, \mathcal{T}) \cdot \mathtt{P}_{AF}(PD|\theta_A)$. A similar prior distribution can be defined by combining the DME and AF prior distributions. Interestingly, mutually exclusive prior distributions (IME and DME) for peptide identification combined with an AF prior distribution for a database will lead to mutually exclusive distributions, because in any specific database $PD$, the probability for two peptides to be identified simultaneously is zero, then the same holds after $PD$ is marginalized out. Specifically, we have the following statement for the AF-DME and AF-IME distributions.

**Proposition 3.3** (AF-DME distribution)**.** *In the* de novo *peptide sequencing algorithm, rather than using a deterministic database $A^*$, we assume an AF prior distribution over databases of size $J$, i.e. $\mathtt{P}_{AF}(PD|\theta_A, \tau_A) = \prod_{j=1}^J \mathtt{P}(\pi_j|\theta_A, \tau_A)$, and for any given peptide database $PD = (\pi_1, \ldots, \pi_J)$, we further assume a DME distribution is used for peptide identification with $l_{PD} = (l_{\pi_j})_{j=1,\ldots,J}$ as parameter, where $l_{\pi_j}$ is independent of the database $PD$, i.e. $\mathtt{P}_{DME}(y_k|PD, \mathcal{T}) = DME(y_k; l_{PD}) \cdot \mathcal{I}(\sum_{\pi \notin PD} y_{k\pi} = 0)$, then $\mathtt{P}(y_k|\theta_A, \tau_A, \mathcal{T})$ follows an DME distribution $DME(y_k; l_{A^*})$ over database $A^*$, with parameters $l_{A^*} = (|A^*| \cdot \mathtt{P}(\pi|\theta_A, \tau_A) \cdot l_\pi)_{\pi \in A^*}$.*

*Proof.* Notice that the peptide database $PD = (\pi_1, \ldots, \pi_J)$ is a vector of random variables, each with $A^*$ as the sample space. Also notice that by using $l_{PD}$ as parameters across different $PD$, we implicitly assume that $l_\pi$ is only dependent on the peptide $\pi$, but not on the peptide database $PD$

[12] This assertion is obviously correct for haploid organisms. For a diploid organism, we may need to consider two distinct peptide sequences, derived from both the major and minor alleles, respectively, and the prior distribution between these two copies can be defined based on the population frequencies of the two alleles.

[13] Note that this assumption can be relaxed after incorporating the estimated protein abundances into the Bayesian model as shown in Figure 4.1a; see Section 4.

containing $\pi$. Hence, we have,

$$\begin{aligned}
&\mathtt{P}_{AF\text{-}DME}(y_k = (y_{k\pi})_{\pi\in A^*}|\theta_A, \tau_A, \mathcal{T}) \\
&\quad = \sum_{|PD|=J,\ PD\subseteq A^*} [\mathtt{P}_{DME}(y_k|PD, \mathcal{T}) \cdot \mathtt{P}_{AF}(PD|\theta_A, \tau_A)] \\
&\quad = \sum_{|PD|=J,\ PD\subseteq A^*} \Bigg\{ \Bigg[ \mathcal{I}\Bigg(\sum_{j=1}^{J} y_{k\pi_j} = 0\Bigg) \cdot \Bigg(1 - \frac{1}{J} \cdot \sum_{j=1}^{J} l_{\pi_j}\Bigg) \\
&\qquad + \frac{1}{J} \cdot \mathcal{I}\Bigg(\sum_{j=1}^{J} y_{k\pi_j} = 1\Bigg) \cdot \prod_{j=1}^{J} l_{\pi_j}^{y_{k\pi_j}} \Bigg] \\
&\qquad \cdot \mathcal{I}\Bigg(\sum_{\pi\notin PD} y_{k\pi} = 0\Bigg) \cdot \prod_{j=1}^{J} \mathtt{P}(\pi_j|\theta_A, \tau_A) \Bigg\} \\
&\quad = \mathcal{I}\Bigg(\sum_{\pi\in A^*} y_{k\pi} = 0\Bigg) \cdot \Bigg(1 - \frac{1}{J} \cdot \mathbb{E}_{PD}\Bigg(\sum_{j=1}^{J} l_{\pi_j}\Bigg)\Bigg) \\
&\qquad + \frac{1}{J} \cdot \mathbb{E}_{PD}\Bigg(\mathcal{I}\Bigg(\sum_{\pi\notin PD} y_{k\pi} = 0\Bigg) \\
&\qquad \cdot \sum_{j=1}^{J}\Bigg(\mathcal{I}\Bigg(\sum_{j=1}^{J} y_{k\pi_j} = 1\Bigg) \cdot l_{\pi_j} \cdot y_{k\pi_j}\Bigg)\Bigg) \\
&\quad = (1 - \mathbb{E}_{\pi_1}(l_{\pi_1})) \cdot \mathcal{I}\Bigg(\sum_{\pi\in A^*} y_{k\pi} = 0\Bigg) \\
&\qquad + \mathbb{E}_{\pi_1}(l_{\pi_1} \cdot \mathcal{I}(y_{k\pi_1} = 1,\ y_{k\pi\neq\pi_1} = 0)) \\
&\quad = \mathcal{I}\Bigg(\sum_{\pi\in A^*} y_{k\pi} = 0\Bigg) \cdot \Bigg(1 - \sum_{\pi\subseteq A^*} l_\pi \cdot \mathtt{P}(\pi|\theta_A, \tau_A)\Bigg) \\
&\qquad + \sum_{\pi\subseteq A^*} [\mathtt{P}(\pi|\theta_A, \tau_A) \cdot l_\pi \cdot \mathcal{I}(y_{k\pi} = 1,\ y_{k\pi'\neq\pi} = 0)] \\
&\quad = \mathcal{I}\Bigg(\sum_{\pi\in A^*} y_{k\pi} = 0\Bigg) \cdot \Bigg(1 - \sum_{\pi\subseteq A^*} l_\pi \cdot \mathtt{P}(\pi|\theta_A, \tau_A)\Bigg) \\
&\qquad + \mathcal{I}\Bigg(\sum_{\pi\in A^*} y_{k\pi} = 1\Bigg) \cdot \sum_{\pi\subseteq A^*} [\mathtt{P}(\pi|\theta_A, \tau_A) \cdot l_\pi]^{y_\pi} \\
&\quad = DME(y_k; l_{A^*}). \qquad \square
\end{aligned}$$

Notice that $\mathtt{P}_{AF\text{-}DME}(y_k|\theta_A, \tau_A, \mathcal{T})$ is independent of the assumed database size $J$, and a nice symmetry is followed by

$$\begin{aligned}
\mathtt{P}_{DME}(y_k|PD, \mathcal{T}) &= \mathcal{I}\Bigg(\sum_{\pi\in PD} y_{k\pi} = 0\Bigg) \cdot (1 - \mathbb{E}_{\pi\in PD}(l_\pi)) \\
&\quad + \mathcal{I}\Bigg(\sum_{\pi\in PD} y_{k\pi} = 1\Bigg) \cdot \prod_{\pi\in PD} \left[\frac{1}{J} \cdot l_\pi\right]^{y_{k\pi}}
\end{aligned}$$

and

$$\begin{aligned}
&\mathtt{P}_{AF\text{-}DME}(y_k|\theta_A, \tau_A, \mathcal{T}) \\
&\quad = \mathcal{I}\Bigg(\sum_{\pi\in A^*} y_{k\pi} = 0\Bigg) \cdot (1 - \mathbb{E}_{\pi\in A^*}(l_\pi)) \\
&\qquad + \mathcal{I}\Bigg(\sum_{\pi\in A^*} y_{k\pi} = 1\Bigg) \cdot \sum_{\pi\subseteq A^*} [\mathtt{P}(\pi|\theta_A, \tau_A) \cdot l_\pi]^{y_{k\pi}},
\end{aligned}$$

with $PD$ replace by $A^*$ and $\frac{1}{J}$ replaced by $\mathtt{P}(\pi|\theta_A, \tau_A)$.

Similar conclusion does not hold for the IID and DID prior distributions. The AF-DID distribution, whose degree is no bigger than $J$, only approximates a DID distribution over $A^*$, whose degree can be $\infty$. For the *de novo* peptide sequencing, we can go one step further and combine AF with mtDME, which again gives a DME distribution. Similarly, for the case of mutant peptide identification, the mutDB prior distribution for databases can be combined with the DME (or mtDME) and obtain the mut-DME, which is still a DME distribution with updated parameters. Therefore, same as for the precursor mass and retention time, the peptide database prior may also be integrated into the prior distribution for peptide identification (equations 3.6 and 3.7) by modifying the peptide detectabilities.

We emphasize that a fixed database $PD_0$ can be interpreted as a specific type of prior distribution, i.e. $\mathcal{I}(PD = PD_0)$, with probability 1 for $PD = PD_0$ and probability 0 for any other database. Thus, the peptide prior distributions (IME, DME, IID and DID) for a given database have already assumed a (simple) database prior distribution. We emphasize that a deterministic database does not make the peptide identification easier, because even when we are certain that proteins/peptides in a database are present in the sample, the peptides may not be detected in a proteomics experiment [5].

### 3.6 Entropy and effective database size of prior distributions

To systematically compare the information content of the prior distributions, we introduce a measure called *effective database size*, defined based on the entropy of an arbitrary prior distribution $\Psi$ for the indicator vector of peptide identification $y$ defined over $\{0,1\}^J$. Let $q_\Psi \triangleq \mathtt{P}_\Psi(y = 0)$ be the *non-matching probability* for $\Psi$. Notice that $J$ and *non-matching probability* $q$ will completely define an IME distribution $IME(1-q, J)$. Let $H_\Psi$ be the entropy of distribution $\Psi$ and $H_{IME}(q, J)$ be the entropy of $IME(1-q, J)$.

**Definition 3.4.** The *effective database size* of the distribution $\Psi$, $\mathcal{N}_\Psi$, is the size of a database on which an IME prior distribution has the same entropy and same *non-matching probability* as that of the distribution $\Psi$, i.e. $\mathcal{N}_\Psi$ satisfies $H_{IME}(q_\Psi, \mathcal{N}_\Psi) = H_\Psi$.

For the IME prior distribution, it is easy to obtain the entropy $H_{IME}(q_\Psi, J) = H(q_\Psi) + (1 - q_\Psi) \cdot \ln N$, where $H(q_\Psi)$ is the entropy for a Bernoulli distribution with parameter $q_\Psi$. Thus, according to the definition, we have

$$\mathcal{N}_\Psi = e^{\frac{H_\Psi - H(q_\Psi)}{1 - q_\Psi}}. \tag{3.19}$$

Note that, for an MS/MS spectrum $s_k$, $\mathcal{N}_\Psi$ is defined for any distribution of $y_k$, including the identification posterior probability $\mathtt{P}(y_k|s_k, \theta, \mathcal{T})$, but not defined for prior distributions over peptide database $\mathtt{P}(PD|\theta_A)$.

With the help from effective database size, we can study the impact of prior information on peptide identification. First, we argue that improving peptide identification is manifested as reducing the effective database size. Apparently, the goal of peptide identification is to assign high (close to 1) probability to the true peptides, $y_k = y_k^{\text{true}}$ for each spectrum $s_k$ when $y_k^{\text{true}} \neq 0$, or assign *non-matching probability* $q_\Psi \to 1$ when $y_k^{\text{true}} = 0$. It is easy to show that when $P(y_k = y_k^{\text{true}}) \to 1$ and $y_k^{\text{true}} \neq 0$, the entropy $H \to 0$ and effective database size $\mathcal{N} \to 1$, which is the minimum value of effective database size; meanwhile when $P(y_k = 0) \to 1$, $H \to 0$ but not necessarily $\mathcal{N} \to 1$. Hence, the process of improving peptide identification is a process of reducing the entropy and effective database size of the distribution $P(y_k)$. However, the opposite statement is not correct, because when the effective database size approaches 1, a wrong peptide may receive the highest probability.

We discussed earlier that a DME or DID prior distribution defined over a larger database could effectively mimic an IME or IID prior distribution defined over a smaller database. In general, it can be proved that $\mathcal{N}_{DME} \leq \mathcal{N}_{IME}$, i.e. the effective database size of a DME distribution is smaller than the original database size, and also $\mathcal{N}_{DID} \leq \mathcal{N}_{IID}$ if $\forall j : d_j < 0.7822$.[14] In practice, it is expected that an informative prior distribution for peptide identification corresponds to a small effective database size, the exact value of which depends on the parameters of the distribution. Finally, we emphasize that the effective database size or entropy are not sufficient for the evaluation of the impact of prior distributions, because even if two prior distributions have the same entropy, one of them could express the *correct* prior information, and hence improve the peptide identification, whereas another one could express *wrong* prior information, and thus deteriorate the peptide identification. More on the appropriateness of prior distributions will be discussed in the discussion section. We note here the appropriateness is hard to quantify unless we have the ground truth of which peptides are the true peptides in the sample.

## 4. THE BAYESIAN APPROACH TO PROTEIN INFERENCE

In previous sections, we focused on the Bayesian model of peptide identification, in which we neglect the mutual dependence among peptides within a same protein and the *degeneracy of* peptides that are shared by multiple proteins. In order to incorporate these aspects, we have proposed a Bayesian model for protein inference that takes as input a set of identified peptides and infers a most probable set of proteins containing these peptides [21, 22]. The Bayesian protein inference model MSBayesPro, as shown in Figure 4.1b, incorporated the (per experiment) peptide detectabilities as prior probabilities, and thus can distinguish proteins sharing the same identified peptides based on the distinct peptides between them that are not identified by search engines. The Bayesian inference models can also be used to improve the peptide identification by computing a posterior probability for each identified peptide [22]. This probability can be viewed as a re-evaluation of the possibility of each peptide identification being correct, taking into account not only the quality of the matching between the peptide and the MS/MS spectrum, but also the mutual dependence among identified peptides (and their matching with the corresponding MS/MS spectra) from the same protein. For example, intuitively, if an identified peptide $a$ from protein $A$ has a high probability of being correct, another identified peptide $b$ from the same protein $A$ will be likely to be correct even if it is scored relatively low by a peptide search engine. It is better to assess the correctness of the peptide identification with the protein inference step if data from a whole LC-MS/MS experiment rather than a single MS/MS spectrum is given.

One technical issue on integrating the MSBayesPro model (Figure 4.1b) into peptide identification is that the protein inference is conducted on the whole experiment rather than each MS/MS spectrum. The peptides need to be identified for each spectrum first and then the spectra corresponding to a single peptide are aggregated, and thus the model can only be applied *after* peptide identification. *Bona fide* Bayesian peptide identification can be achieved by the model proposed in Figure 4.1a, which is similar to the model in the Figure 3.1d, but explicitly model the protein sample $P$ as protein sequence and abundance pairs $(x_i, q_i)_{i=1,\ldots,I}$ rather than a peptide database $PD = (\pi_j)_{j=1,\ldots,J}$ with peptides of equal abundance. By considering the information from the other MS/MS spectra, the model can effectively estimate a more specific prior distribution (and hence smaller effective database size) for each single MS/MS spectrum.

The Bayesian approach can be further extended to address several open problems in proteomics. As shown in Figure 4.1c, after inserting an extra layer of nodes (denoted as $u$) in the Bayesian model, it is possible to infer the expression of a *protein set*[15] in the sample, such as co-regulated genes, protein complexes and proteins in the same pathway. The model can also be applied to the analysis of metaproteomics data in principle to infer the presence of species

[14] This can be shown via the convexity of $g(d_j) = -(1 - e^{\ln(1-d_j)}) \cdot [\ln(1 - d_j) - \ln(1 - e^{\ln(1-d_j)})]$ when $d_j \in (0, 0.7822)$. Note that $d_j$ is the per spectra peptide detectability, thus the values of $d_j$ should be at the order of $1/J$. Hence,the condition $d_j < 0.7822$ is expected to be true.

[15] We use *protein set* to denote a group of proteins expected to have positively correlated expression levels. A related notation, the *gene set*, is commonly used in gene expression analysis [32].

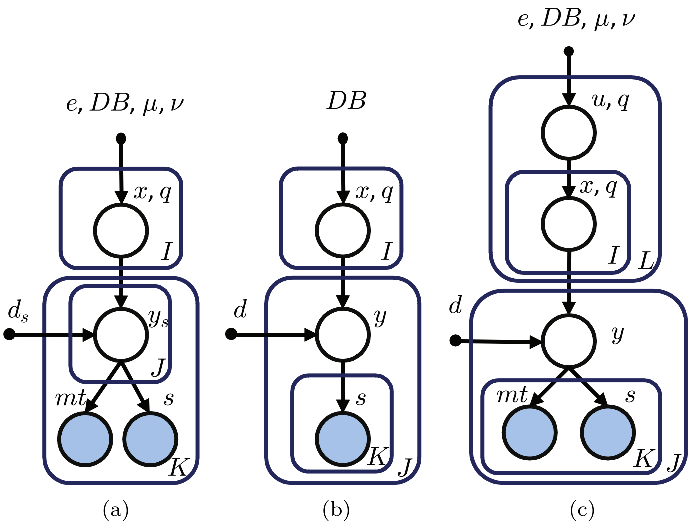

*Figure 4.1. Bayesian probabilistic models considering protein or higher level information for the LC-MS/MS analysis of proteome samples. (a) A Bayesian model for peptide identification incorporating prior information at protein level; (b) The MSBayesPro [22] model for protein inference; (c) A Bayesian model for peptide and protein identification incorporating prior information from protein sets (e.g. protein complexes or proteins in the same pathways). Graphical notations in addition to those in Figure 3.1: $x$, random variable for the presence of a protein in a sample; $q$: random variable for the abundance of a protein; $u$,* protein set *such as the set of proteins from a gene, a protein complex, a signaling pathway, or proteins from one organism in the context of meta-proteomics; $I$, $L$, the numbers of proteins and protein sets considered in an analysis.*

bases on proteomics evidences. Extensive further studies are needed to explore the applications of such a model.

We want to emphasize here that the models in Figure 3.1b to 4.1a will be able to make use of incrementally more prior information, and hence should be able to provide incrementally better performances for the same tasks. For example, the model in Figure 4.1a is expected to give better peptide identification than the model in Figure 3.1c. One implication is that the *de novo* peptide sequencing algorithms coupled (in a post processing fashion) with the advanced models for the identification of protein or *protein set* might be able to achieve a good power for confidently identifying peptides, even if it cannot be done with model 3.1d. This will also be an interesting problem for future investigation.

## 4.1 Using protein abundances as prior information

For most prior distributions discussed so far, the peptide and protein abundance information are not considered. This simplification may not always be appropriate. For example, unlike mutations in peptides which either occur or not, post-translational modification processes are quantitative and dynamic. Furthermore, the modified and the original peptides may be both present in the sample at various abundance ratios. To incorporate abundance information, we need 1) to specify prior distribution for peptide abundances; and 2) to model the dependence of peptide identification on the peptide quantity in an LC-MS/MS analysis, similar as the database priors and identification priors that we have discussed in previous sections. The simplest peptide abundance prior adopts a delta distribution $\delta(q_j - q_{j_0})$, i.e. to treat $q_j$ as a deterministic parameter with unknown value $q_{j_0}$. The quantity dependent peptide identification prior can be devised by using the quantity adjustment of peptide detectability (resulting in *effective detectability*, as used in the MSBayesPro model [23, 21], Figure 4.1b) via the transformation $P(y_j = 1|q_j, DB, \mathcal{T}) = 1 - (1 - d_j)^{q_j}$ for each peptide $j$. These can be implemented for the model of Figure 4.1a to improve Bayesian peptide identification.

The dilemma of using protein abundance as prior information is that they are not available until peptide identification and protein inference are completed. This, however, does not prevent using an iterative procedure to improve peptide and protein identification *a posteriori*. In order to analyze the effect of considering protein abundances in peptide identification, we assume that the abundances of all proteins in a sample are precisely known, whether the proteins are identified or not, and we assume proteins do not share proteolytic peptides. These are not realistic assumptions in the current protocol of proteomics data analysis. They are assumed here for simplicity so that peptide identification is de-coupled from protein inference. Under these assumptions, we argue (instead of proving) that knowing the protein abundances should *on average* lead to a smaller entropy for the DME distribution for peptide identification than assuming all proteins to be at equal abundance. First, *on average*, the mean standard peptide detectability $\bar{d}_i$ for peptides in each protein $i$ is roughly a constant. If we further assume all proteins (roughly) have the same number of peptides, then we can use an *average peptide* with detectability $\bar{d}_i$ to represent all peptides in each protein. As a result, assuming equal protein abundances will lead to an IME prior distribution over *average peptides*, whereas knowing the real protein abundances will lead to a DME distribution and hence a smaller effective database size.

# 5. DISCUSSION

Incorporating prior knowledge to reduce the search space and to improve protein identification is a common strategy applied in current proteomics research. A Bayesian probabilistic point of view of this strategy, however, has not been presented before. The Bayesian framework for protein identification is meaningful in three ways. First, it provides a formal yet intuitive framework for understanding the advantages and pitfalls of the current protein identification algorithms. Second, as a robust and extendible framework, the Bayesian model enables further improvement of the current algorithms and the development of new algorithms to address challenges yet to be solved, e.g. the identification of mutations and PTMs, based on probabilistic principles. Finally, the Bayesian framework provides probabilistic scores that can be intuitively interpreted for peptide and protein identification. However, there are several issues with the Bayesian framework, such as the potentially high computational complexity and the required prior information, which we discuss below.

## 5.1 Computational complexity of the Bayesian model

The Bayesian model for peptide identification can either be used as a post-processing step after database searching, or integrated into the peptide identification algorithms. In principle, in a Bayesian peptide identification framework, all peptides with non-zero prior probabilities need to be matched with MS/MS spectra explicitly or implicitly. This requires extensive computing when the database is large. In practice, however, sophisticated algorithms such as dynamic programing are used in the *de novo* peptide sequencing to avoid brute force peptide-spectrum matching over all peptides.

An alternative strategy is to filter out all peptides $j$ with sufficiently small prior probability $\mathsf{P}(y_{kj} = 1|DB, \mathcal{T}) < \epsilon_0$, and reduce the search space. A similar strategy was typically applied in the precursor mass based filtering of peptides used in the peptide search engines, and sometimes was integrated into the PSM scoring scheme [31]. Notably, in the prior probability filtering in the Bayesian framework, the peptides passing the filtering will still be assigned different prior probabilities, while peptides passing a precursor mass based filter will effectively be assigned with equal probabilities in non-probabilistic peptide identification algorithms.

The prior probability filtering strategy may accelerate peptide identification algorithms; in particular, when a proper and informative prior is available. For example, we may filter the peptide database $PD$ for a spectrum $s_k$ by the precursor mass and retention time modified peptide detectability $d_{kj} = \mathsf{P}(y_{kj} = 1|m_k, t_k, PD, \mathcal{T})$ for each candidate peptide $\pi_j$, where $m_k$ and $t_k$ are the observed precursor mass and retention time for $s_k$. Compared with the precursor-mass-based filter used in database searching algorithms, this filter will result in an even smaller number of peptides to be matched to each MS/MS spectrum.

## 5.2 Appropriate prior distributions

The application of the prior distributions in the Bayesian model for peptide and protein identification is dependent on the prior knowledge of the proteome sample. For an organism without extensive previous studies, a proteomics study may have to rely mainly on *de novo* peptide sequencing, in which only the prior information regarding the experimental platform and protocol (e.g. fragmentation ion types and preferences, mass accuracy, and peptide detectability which summarized multiple factors) can be utilized. If a protein sequence database from related species are available, error-tolerant database searching (which effectively enlarge the reference database with all possible mutations, or in our Bayesian framework, assign non-zero prior probabilities to mutant peptides) could be used. When the genomic or transcriptomic sequences of an organism are available, the derived protein sequence database can be used by database search engines for peptide identification. With the accumulation of more genomic, transcriptomic and proteomic studies carried out on a specific type of sample (e.g. a specific tissue or a specific development stage of a specific organism), informative prior information about the presence of proteins and their abundances can be used in subsequent proteomics studies of the same type of sample. Ultimately, we may know for sure the protein sequences and their abundances for a given type of sample (and thus no further protein identifications are needed), we may still use the prior information for the identification of mutations or PTMs.

To facilitate the use of the prior information gained from previous experimental studies, it is important to construct the repository of these data. Compared with genomic and transcriptomic data, proteomic databases are less well developed, partly due to lack of standardization of current proteomics techniques, the complexity of proteomics datasets, and the lack of standard proteomics tool sets. Most of the current proteomics data repositories [10, 12, 36] collect raw proteomics data. The GPMDB (Global Proteome Machine Database) [9] is a unique database for hosting processed proteomics data. Not without limitations, GPMDB serves as a good model for the proteomics knowledge database that are useful both for Bayesian protein identification of future proteomics experiment and for integrative studies across multiple data sources – genomics, transcriptomics, and proteomics.

The advantages of using informative prior distribution for protein identification come with a price. A prior probability distribution is appropriate only when it expresses the correct prior knowledge regarding a proteome sample. Just like using the protein sequence database from an irrelevant species, using an inappropriate prior distribution can lead to poor performance of peptide identification. For example, while accurate peptide detectabilities used for peptide identification can improve protein inference [22, 23], using peptide detectability trained on a different experimental platform may result in worse protein inference. A specific (low entropy) but inappropriate prior can be worse than a non-informative prior by biasing the posterior probabilities incorrectly. As a result, it is critical to select informative and appropriate prior distribution for the specific datasets to be analyzed.

## ACKNOWLEDGEMENTS

The authors would like to thank Dr. Karen Kafadar for comments on the manuscript, Dr. Nuno Bandeira for discussions on the relationship between peptide detectability and LC retention time, Drs. Pavel Pevzner and Nitin Gupta for discussions on applying peptide detectability in the identification of peptides containing mutations. This work is partially supported by NIH grant R01 RR024236-01A1.

*Received 5 May 2011*

## REFERENCES

[1] Alves, P., Arnold, R. J., Clemmer, D. E., Li, Y., Reilly, J. P., Sheng, Q., Tang, H., Xun, Z., Zeng, R., and Radivojac, P. Fast and accurate identification of semi-tryptic peptides in shotgun proteomics. *Bioinformatics 24*, 1 (Jan 2008), 102–109.

[2] Alves, P., Arnold, R. J., Novotny, M. V., Radivojac, P., Reilly, J. P., and Tang, H. Advancement in protein inference from shotgun proteomics using peptide detectability. *Pac Symp Biocomput* (2007), 409–420.

[3] Beausoleil, S. A., Jedrychowski, M., Schwartz, D., Elias, J. E., Villén, J., Li, J., Cohn, M. A., Cantley, L. C., and Gygi, S. P. Large-scale characterization of hela cell nuclear phosphoproteins. *Proc Natl Acad Sci USA 101*, 33 (Aug 2004), 12130–12135.

[4] Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J., and Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol 24*, 10 (Oct 2006), 1285–1292.

[5] Bohrer, B. C., Li, Y. F., Reilly, J. P., Clemmer, D. E., DiMarchi, R. D., Radivojac, P., Tang, H., and Arnold, R. J. Combinatorial libraries of synthetic peptides as a model for shotgun proteomics. *Anal Chem 82*, 15 (Aug 2010), 6559–6568.

[6] Chen, X., Drogaris, P., and Bern, M. Identification of tandem mass spectra of mixtures of isomeric peptides. *J Proteome Res 9*, 6 (Jun 2010), 3270–3279.

[7] Consortium, U. The universal protein resource (uniprot) in 2010. *Nucleic Acids Res 38*, Database issue (Jan 2010), D142–D148.

[8] Craig, R. and Beavis, R. C. Tandem: matching proteins with tandem mass spectra. *Bioinformatics 20*, 9 (Jun 2004), 1466–1467.

[9] Craig, R. and Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics 20*, 9 (Jun 2004), 1466–1467.

[10] Deutsch, E. W., Lam, H., and Aebersold, R. Peptideatlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep 9*, 5 (May 2008), 429–434.

[11] Elias, J. E. and Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods 4*, 3 (Mar 2007), 207–214.

[12] Falkner, J. A., Falkner, J. W., and Andrews, P. C. Proteomecommons.org jaf: reference information and tools for proteomics. *Bioinformatics 22*, 5 (Mar 2006), 632–633.

[13] Fenyö, D. and Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem 75*, 4 (Feb 2003), 768–774.

[14] Frank, A. and Pevzner, P. Pepnovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem 77*, 4 (Feb 2005), 964–973.

[15] Gauci, S., Helbig, A. O., Slijper, M., Krijgsveld, J., Heck, A. J. R., and Mohammed, S. Lys-n and trypsin cover complementary parts of the phosphoproteome in a refined scx-based approach. *Anal Chem 81*, 11 (Jun 2009), 4493–4501.

[16] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. Open mass spectrometry search algorithm. *J Proteome Res 3*, 5 (2004), 958–964.

[17] Henikoff, S. and Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA 89*, 22 (Nov 1992), 10915–10919.

[18] Juncker, A. S., Jensen, L. J., Pierleoni, A., Bernsel, A., Tress, M. L., Bork, P., von Heijne, G., Valencia, A., Ouzounis, C. A., Casadio, R., and Brunak, S. Sequence-based feature prediction and annotation of proteins. *Genome Biol 10*, 2 (2009), 206.

[19] Keller, A., Nesvizhskii, A., Kolker, E., and Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Anal Chem 74*, 20 (2002), 5383–5392.

[20] Kim, S., Gupta, N., and Pevzner, P. A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res 7*, 8 (Aug 2008), 3354–3363.

[21] Li, Y. F., Arnold, R. J., Li, Y., Radivojac, P., Sheng, Q., and Tang, H. A Bayesian Approach to Protein Inference Problem in Shotgun Proteomics. *RECOMB (Research in Computational Molecular Biology)* (2008), 167–180.

[22] Li, Y. F., Arnold, R. J., Li, Y., Radivojac, P., Sheng, Q., and Tang, H. A bayesian approach to protein inference problem in shotgun proteomics. *J Comput Biol 16*, 8 (Aug 2009), 1183–1193. MR2534085

[23] Li, Y. F., Arnold, R. J., Tang, H., and Radivojac, P. The importance of peptide detectability for protein identification, quantification, and experiment design in MS/MS proteomics. *Journal of Proteome Research 9*, 12 (2010), 6288–6297.

[24] Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom 17*, 20 (2003), 2337–2342.

[25] Ma, Z.-Q., Dasari, S., Chambers, M. C., Litton, M. D., Sobecki, S. M., Zimmerman, L. J., Halvey, P. J., Schilling, B., Drake, P. M., Gibson, B. W., and Tabb, D. L. Idpicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res 8*, 8 (Aug 2009), 3872–3881.

[26] Nesvizhskii, A., Keller, A., Kolker, E., and Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem 75*, 17 (2003), 4646–4658.

[27] Nesvizhskii, A. I. and Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics 4*, 10 (Oct 2005), 1419–1440.

[28] Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis 20*, 18 (Dec 1999), 3551–3567.

[29] Reiter, L., Claassen, M., Schrimpf, S. P., Jovanovic, M., Schmidt, A., Buhmann, J. M., Hengartner, M. O., and Aebersold, R. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol Cell Proteomics 8*, 11 (2009), 2405–17.

[30] Serang, O., MacCoss, M. J., and Noble, W. S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *Journal of Proteome Research 9*, 10 (2010), 5346–5357. doi: 10.1021/pr100594k.

[31] Strittmatter, E. F., Kangas, L. J., Petritis, K., Mottaz, H. M., Anderson, G. A., Shen, Y., Jacobs, J. M., Camp, D. G., and Smith, R. D. Application of peptide lc retention time information in a discriminant function for peptide identification by tandem mass spectrometry. *J Proteome Res 3*, 4 (2004), 760–769.

[32] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA 102*, 43 (Oct 2005), 15545–15550.

[33] Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M., Ham, A.-J. L., Bunk, D. M., Kilpatrick, L. E., Billheimer, D. D., Blackman, R. K., Cardasis, H. L., Carr, S. A., Clauser, K. R., Jaffe, J. D., Kowalski, K. A., Neubert, T. A., Regnier, F. E., Schilling, B., Tegeler, T. J., Wang, M., Wang, P., Whiteaker, J. R., Zimmerman, L. J., Fisher, S. J., Gibson, B. W., Kinsinger, C. R., Mesri, M., Rodriguez, H., Stein, S. E., Tempst, P., Paulovich, A. G., Liebler, D. C., and Spiegelman, C. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J Proteome Res* (Dec 2009).

[34] Tang, H., Arnold, R. J., Alves, P., Xun, Z., Clemmer, D. E., Novotny, M. V., Reilly, J. P., and Radivojac, P. A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics 22*, 14 (Jul 2006), e481–e488.

[35] Tanner, S., Shu, H., Frank, A., Wang, L.-C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. Inspect: Identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem 77*, 14 (Jul 2005), 4626–4639.

[36] Vizcaíno, J. A., Côté, R., Reisinger, F., Foster, J. M., Mueller, M., Rameseder, J., Hermjakob, H., and Martens, L. A guide to the proteomics identifications database proteomics data repository. *Proteomics 9*, 18 (Sep 2009), 4276–4283.

[37] Wan, Y., Yang, A., and Chen, T. Pephmm: A hidden markov model based scoring function for mass spectrometry database search. *Anal Chem 78*, 2 (Jan 2006), 432–437.

[38] Wang, Z., Gerstein, M., and Snyder, M. Rna-seq: A revolutionary tool for transcriptomics. *Nat Rev Genet 10*, 1 (Jan 2009), 57–63.

[39] Weatherly, D. B., J. A. Atwood, r., Minning, T. A., Cavola, C., Tarleton, R. L., and Orlando, R. A heuristic method for assigning a false-discovery rate for protein identifications from mascot database search results. *Mol Cell Proteomics 4*, 6 (2005), 762–72.

[40] Yates, J. R., Eng, J. K., McCormack, A. L., and Schieltz, D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem 67*, 8 (Apr 1995), 1426–1436.

[41] Zhang, B., Chambers, M. C., and Tabb, D. L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J Proteome Res 6*, 9 (Sep 2007), 3549–3557.

[42] Zhang, N., jun Li, X., Ye, M., Pan, S., Schwikowski, B., and Aebersold, R. Probidtree: An automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics 5*, 16 (Nov 2005), 4096–4106.

Yong Fuga Li
School of Informatics and Computing
Indiana University
Bloomington, IN 47405
USA
E-mail address: yonli@indiana.edu

Randy J. Arnold
Department of Chemistry
Indiana University
Bloomington, IN 47406
USA
E-mail address: rarnold@indiana.edu

Predrag Radivojac
School of Informatics and Computing
Indiana University
Bloomington, IN 47405
USA
E-mail address: predrag@indiana.edu

Haixu Tang
School of Informatics and Computing
Indiana University
Bloomington, IN 47405
USA
E-mail address: hatang@indiana.edu