

Proteome Coverage Prediction for Integrated Proteomics Datasets

MANFRED CLAASSEN,^{1,2,3} RUEDI AEBERSOLD,³ and JOACHIM M. BUHMANN¹

ABSTRACT

Comprehensive characterization of a proteome defines a fundamental goal in proteomics. In order to maximize proteome coverage for a complex protein mixture, i.e., to identify as many proteins as possible, various different fractionation experiments are typically performed and the individual fractions are subjected to mass spectrometric analysis. The resulting data are integrated into large and heterogeneous datasets. Proteome coverage prediction refers to the task of extrapolating the number of protein discoveries by future measurements conditioned on a sequence of already performed measurements. Proteome coverage prediction at an early stage enables experimentalists to design and plan efficient proteomics studies. To date, there does not exist any method that reliably predicts proteome coverage from integrated datasets. We present a generalized hierarchical Pitman-Yor process model that explicitly captures the redundancy within integrated datasets. The accuracy of our approach for proteome coverage prediction is assessed by applying it to an integrated proteomics dataset for the bacterium *L. interrogans*. The proposed procedure outperforms ad hoc extrapolation methods and prediction methods designed for non-integrated datasets. Furthermore, the maximally achievable proteome coverage is estimated for the experimental setup underlying the *L. interrogans* dataset. We discuss the implications of our results for determining rational stop criteria and their influence on the design of efficient and reliable proteomics studies.

Key words: algorithms, computational molecular biology.

1. INTRODUCTION

RECENT DEVELOPMENTS IN MASS SPECTROMETRY-BASED PROTEOMICS have enabled biologists to comprehensively characterize proteomes, the protein inventories of biological samples (Domon and Aebersold, 2006). To achieve extensive proteome coverage, a range of different experiments have to be carefully planned and extensively repeated. *Proteome coverage prediction* denotes the task of estimating the expected yield of protein discoveries upon repetitions of experiments. This task is essential in the context of the more general task of experimental planning, aiming to infer a particular series of experiments that achieves maximal coverage. Here we present a generalized hierarchical Pitman-Yor process to reliably predict proteome coverage for multidimensional fractionation experiments.

¹Department of Computer Science and ²Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland.
³Center for Systems Physiology and Metabolic Diseases, Zurich, Switzerland.

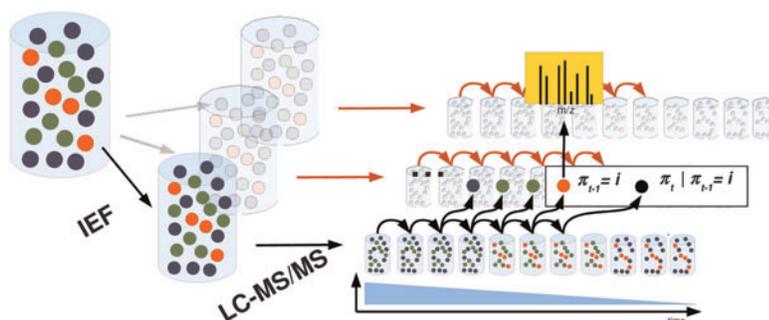
The most successful strategy to achieve extensive proteome coverage is referred to as shotgun proteomics. Briefly, proteins are biochemically extracted from a biological sample and are enzymatically digested to yield a complex ensemble of peptides. Protein and/or peptide ensembles are optionally further fractionated according to physical/chemical/biological properties (multidimensional fractionation). Tandem mass spectrometry is then employed to sample and identify individual peptide species present in the resulting ensembles, and to finally recover the set of proteins initially present in the biological sample (Nesvizhskii et al., 2007) (Fig. 1).

The capacity of mass spectrometers limits the number of peptides possibly identified at a time. Due to this constraint, it is by far too difficult to identify the entirety of species in a peptide ensemble arising after enzymatic digestion of a typical complex biological sample such as a complete proteome. Two experimental routes are pursued to circumvent this limitation and to enable comprehensive characterization of a complex peptide ensemble. First, peptide ensembles are fractionated into a multitude of less complex and, therefore, more tractable ensembles before being analyzed by tandem mass spectrometry and second, experiments are extensively repeated. Popular fractionation schemes separate peptides with respect to properties such as, for example, size or isoelectric point. Reversed-phase liquid chromatography (LC) as the most common fractionation technique separates peptide ensembles according to hydrophobicity and is typically directly coupled to a tandem mass spectrometry system (LC-MS/MS). Multidimensional fractionation strategies comprise multiple steps of fractionation, typically fractionation according to some physico-chemical property other than hydrophobicity followed by LC-MS/MS analysis (Fig. 1). Shotgun proteomics studies that achieved significant proteome coverage for a variety of organisms have shown to build on extensive repetition of multidimensional fractionation experiments (Brunner et al., 2007).

Methods for proteome coverage prediction estimate the expected number of peptide/protein discoveries when experiments are repeated. The ability to predict proteome coverage provides essential information for rational experimental planning of shotgun proteomics studies. Projects aiming at extensive proteome coverage require a considerable amount of experimentation. Proteome coverage should ideally increase with consecutive experiments in an efficient way. The choice between competing experimental setups has to be guided by their potential to increase proteome coverage. Methods for proteome coverage prediction enable to rationally determine the optimal setup. Proteome coverage prediction furthermore enables to estimate the maximal coverage as well the volume of experiments required to achieve this coverage.

Proteome coverage prediction and related tasks have not been addressed until recently. Eriksson and Fenyo (2007) conducted simulation studies to generally investigate how fractionation of peptide or protein ensembles might affect the efficiency of shotgun proteomics experiments. Brunner *et al.* (2007) roughly estimated upper and lower bounds for proteome coverage from a real data set by assuming worst/best case scenarios. Recently, an infinite Markov model based on Dirichlet processes (Beal et al., 2002) has been proposed to characterize LC-MS/MS experiments and for the first time to predict proteome coverage for one-dimensional fractionation experiments (Claassen et al., 2009).

FIG. 1. Typical multidimensional fractionation experiment. The initial root peptide ensemble obtained from the biological source is separated by some fractionation method (e.g., isoelectric focussing [IEF]), giving rise to a set of related peptide ensembles. LC-MS/MS analysis is performed for each of these fractions. Liquid chromatography fractionation generates a sequence of child peptide ensembles from the root ensemble.



Each of these ensembles is derived from the root ensemble by pooling peptides of similar polarity. The sequence of ensembles features descending overall polarity in the course of the experiment. During the experiment, peptides π_t are drawn from the sequence of ensembles and analyzed by the mass spectrometer coupled to the liquid chromatography system and subsequently identified computationally. We propose a non-parametric Bayesian approach to characterize the distributions governing the peptide ensembles. We simulate further experiments and thereby predict proteome coverage by sampling from these peptide distributions.

In practice, it is highly desirable to predict proteome coverage of multidimensional fractionation experiments since these strategies have shown to have the largest potential to map out a proteome. However, there does not exist any method for proteome coverage prediction of these experiments. This task is particularly challenging since the proteomes represented by each fraction overlap to an unknown extent. Proteome coverage prediction methods for multidimensional fractionation experiments have to account for this phenomenon.

This article generalizes the non-parametric approach to characterize peptide distributions arising in LC-MS/MS experiments (Claassen et al., 2009) to further enable proteome coverage prediction from integrated datasets compiled from multidimensional fractionation experiments. Specifically, we propose a novel generalized hierarchical Pitman-Yor process (Teh et al., 2006; Teh, 2006) with self-referential base measures that addresses the issue of distribution overlap which is introduced by the fractionation preceding the LC-MS/MS analysis. Besides the possibility to characterize peptide distributions arising in the course of multidimensional fractionation experiments, this approach also lends itself to characterize the biologically more relevant protein distributions. We assess our method on a set of 24 experiments from multidimensional fractionation of a *L. interrogans* whole proteome sample and report better performance than ad hoc extrapolation schemes and other approaches designed for one dimensional fractionation experiments. We discuss our results with respect to maximally achievable proteome coverage from a peptide—as well as protein-centric perspective.

2. METHODS

The following sections give technical background and details on the hierarchical Pitman-Yor process framework for proteome coverage prediction based on integrated datasets. Briefly, our approach characterizes the peptide/protein distributions arising in a multidimensional fractionation experiment and simulates further experiments by sampling from these distributions. Proteome coverage is predicted by counting the number of novel peptide/protein discoveries in the simulations. In the following sections, we will assume a peptide-centric view for clarity, i.e., consider peptide distributions instead of its protein counterparts. Note that peptides, by virtue of being protein fragments, also refer to protein identities. Therefore, the following sections can also be read by consequently substituting peptides with proteins. Complications arising from peptides ambiguously referring to several protein identities are discussed in Section 4.

2.1. Pitman-Yor processes

We apply Pitman-Yor processes to characterize peptide distributions arising in the course of a series of proteomics experiments. The following section briefly reviews the concept of Pitman-Yor processes in the context of this work.

Like the Gaussian distribution is an appropriate distribution for a real valued random variable in numerous applications, the Pitman-Yor process prior frequently is an appropriate distribution for complex objects such as discrete distributions (Pitman and Yor, 1997). Informally, Pitman-Yor processes priors are suited as distributions over discrete distributions that are expected to have most of their probability mass on a small number of atoms and only little probability mass on the vast majority of atoms (Teh, 2006). As various proteomics studies have shown that protein/peptide frequencies exhibit such a property (Reiter et al., 2009), we use Pitman-Yor processes priors to characterize distributions G over a set Π of peptides defined by a protein database of the studied organism.

$$G \mid \gamma, d, H \sim \text{PY}(\gamma, d, H) \quad (1)$$

where $\text{PY}(\gamma, d, H)$ is a Pitman-Yor process with a concentration parameter γ , a discount parameter d and a base probability measure H . The base measure is defined over Π (sample space). H is frequently chosen uniform, assigning $1/|\Pi|$ probability mass to each $\pi \in \Pi$.

The so-called *Chinese restaurant* construction (Blackwell and MacQueen, 1973; Pitman, 2002) provides an intuitive way to see which kind of distributions are likely to be drawn from a Pitman-Yor process prior $\text{PY}(\gamma, d, H)$. Imagine a restaurant with an infinite number of tables. At each table, a specific dish is served. The distribution G over dishes is constructed after having seated an infinite number of customers. Customers are seated according to a probabilistic rule. Specifically, the probability of the t -th customer being seated at the table serving dish $\pi_t = k$ assumes the values

$$P(\pi_t = k \mid \pi_1, \dots, \pi_{t-1}, \gamma, d, H) = \begin{cases} \frac{n_k - d}{t-1+\gamma} & \text{populated table} \\ \frac{\gamma + kd}{t-1+\gamma} & \text{next unpopulated table} \end{cases} \quad (2)$$

where n_k corresponds to the number of customers already sitting at the table serving dish i . In case a customer happens to be seated at a new table, the dish served at this table is drawn from the base probability measure H . A procedural description of serving a new customer in a restaurant with seating arrangement $R = n_1, n_2 \dots$ is as follows:

SEAT(R, γ, d, H)

```

1   $t \leftarrow \text{SAMPLETABLE}(R, \gamma, d)$ 
2  if  $t \neq \text{new}$ 
3    then return DISH( $R, t$ )
4    else return SAMPLE( $H$ )

```

The larger the concentration parameter γ , the higher the chances that a new customer is seated at a new table. The more customers have already been seated, the less likely a new dish will be served. The larger the discount parameter d the less likely a customer is seated at an already populated table. Note that $d < 1$. In summary, the parameters γ and d control, though in different ways, the deviation of G from the base measure H . The *Chinese restaurant* construction specifies the posterior to iteratively sample from $\pi_t \mid \pi_1, \dots, \pi_{t-1}, \gamma, d, H$ after marginalizing out G .

Pitman-Yor processes are generalizations of the more commonly known Dirichlet processes (Blackwell and MacQueen, 1973; Antoniak, 1974). More precisely, a Dirichlet process $\text{DP}(\gamma, H)$ is equivalent to a Pitman-Yor process $\text{PY}(\gamma, d, H)$ with $d = 0$. Both Dirichlet and Pitman-Yor processes priors will be used as priors for peptide distributions that arise in the course of a multidimensional fractionation experiment. After having estimated the process parameters, further experiments are simulated by sampling according to the *Chinese restaurant* construction.

2.2. Hierarchical process model for multidimensional fractionation experiments

This section characterizes the distributions which arise in a multidimensional fractionation experiment. We specifically describe a typical setup that comprises two consecutive fractionation steps, where the first step splits the initial peptide ensemble into a set of I fractions that are each analyzed by LC-MS/MS (Fig. 1). Besides enforcing consistency along subsequent fractionation steps using hierarchical processes, our model explicitly captures the similarity of corresponding peptide distributions across different fractions.

The initial peptide ensemble follows the root distribution G . We assume a Pitman-Yor process prior $\text{PY}(\gamma_r, d_r, H)$ for G . The base measure H is chosen to be the uniform distribution over the peptides defined by the protein database of the studied organism.

Peptides are not directly sampled from the root distribution G . Consider some time point t during the LC-MS/MS analysis of fraction i . The peptide π_t^i is sampled from the child peptide distribution G_t^i of the peptide ensemble currently eluting from the liquid chromatography column. Following Claassen et al. (2009), we assume that the precedent peptide $\pi_{t-1}^i := j$ is indicative for the current polarity of the chromatography and thereby the current peptide distribution, i.e., with a slight abuse of notation we assume $G_t^i = G_j^i$. Further, we assume a Dirichlet process prior for G_j^i , resulting in an infinite Markov model for LC-MS/MS experiments similar to Claassen et al. (2009).

$$\begin{aligned} G_j^i \mid \gamma_c^i, A_j^i &\sim \text{DP}(\gamma_c^i, A_j^i) \\ \pi_t \mid \pi_{t-1}^i = j &\sim G_j^i \end{aligned} \quad (3)$$

We want the child distributions G_j^i to be consistent with the root distribution G , i.e., we want to ensure that peptides having zero probability mass in the initial peptide ensemble still have zero probability mass during an LC-MS/MS experiment. This notion is captured by choosing G as base measure A_j^i in (3), which results in a hierarchical process (Teh et al., 2006). This choice ensures (1) that G_j^i is consistent with G , i.e., the support of G_j^i is enclosed by the support of G and (2) that G_j^i will have similarity to G to an extent defined by the concentration parameter γ_c^i . Furthermore, we want to capture the similarity between G_j^i and its

corresponding distributions G_j^i in all other fractions $i' \neq i$. Therefore, we extend the base measure A_j^i in (3) to a (self-referential) linear combination of the distributions $(G_j^i)_{i'=1}^I$ and G .

$$A_j^i = a_i^i G + \sum_{i' \neq i} a_i^{i'} G_j^{i'} \quad (4)$$

Since the values $a^i := (a_{i'}^i)_{i'=1}^I$ are not known beforehand, it is natural to treat them as a random discrete distribution with a Dirichlet process prior. The a_i^i reflect the dissimilarity of fraction i from the other fractions by controlling the rate of sampling peptides directly from the root distribution G . We account for their distinguished role by putting prior weight α_a^i on a_i^i and incorporating this parameter by assuming for the a^i a biased (in the sense of Claassen et al., 2009) Dirichlet process prior $\text{DP}_i(\gamma_a^i, \alpha_a^i, M)$ with uniform base measure $M := (1/I)_{1..J}$. In the following, we will refer to the a^i as the adapter distributions.

The self-referential base measures A_j^i are a crucial component of this process since they capture the important overlap of peptide distributions across the fractions j arising in a multidimensional fractionation experiment. The step from the simple base measure G as described in Claassen et al. (2009) to the self-referential base measure enables to appropriately characterize the peptide distributions describing such an experiment.

Putting together the precedent considerations, the stochastic source of a in a multidimensional fractionation experiment is fully characterized by

$$\begin{aligned} G \mid \gamma_r, d_r, H &\sim \text{PY}(\gamma_r, d_r, H) \\ a^i \mid \gamma_a^i, \alpha_a^i, M &\sim \text{DP}_i(\gamma_a^i, \alpha_a^i, M) \\ G_j^i \mid \gamma_c^i, A_j^i &\sim \text{DP}(\gamma_c^i, A_j^i) \\ \pi_t \mid \pi_{t-1} = j &\sim G_j^i \end{aligned} \quad (5)$$

Note that it is straightforward to assume Pitman-Yor process priors for all distributions. This choice though comes at the cost of additional parameters that have to be learned from data. In this work, we focused on robustness and therefore decided to keep the priors of the child distributions as simple as possible.

2.3. Sampling sequences of protein identifications

This section describes a nested, recursive *Chinese restaurant* construction to sample peptides from the hierarchical process model with self-referential base measures given an already observed series π of already observed peptides, i.e., how to simulate further experiments.

Each distribution in the hierarchical process model has a restaurant representation, i.e., a seating arrangement. Specifically, we denote the restaurants corresponding to the G_j^i as $R_{ij}^c = (n_{ijk}^c)_{k=1}^K$, those to the a^i as $R_i^a = (n_{ii'}^a)_{i'=1}^I$ and the root restaurant as $R^r = (n_k)_{k=1}^K$. To keep the notation uncluttered, we incorporate the prior weights α_a^i into the counts $n_{ii'}^a$ and respectively R_i^a . \mathbf{R} denotes the set of all restaurants. Note that a set of seating arrangements \mathbf{R} implies a series π of observed identifications. We further summarize the set of parameters by $\theta := (\gamma_r, d_r, \gamma_a^1, \dots, \gamma_a^I, \gamma_c^1, \dots, \gamma_c^I)$.

For a given set of seating arrangements \mathbf{R} we now sample the identification π_t for fraction i and preceding identification $\pi_{t-1} = j$. Verbally, we first iterate the *Chinese restaurant* construction for the corresponding child distribution. In case this iteration triggers a sampling event of its base measure, we have to determine which of its mixture components is to be sampled. Therefore, the *Chinese restaurant* construction of the corresponding adapter distribution is iterated. Subsequently, either the root restaurant or, recursively, some of the sibling child restaurants of another fraction is iterated. This procedure can be summarized as shown below.

SAMPLEIDENTIFICATION($i, j, \mathbf{R}, \theta, H, M$)

```

1   $\pi \leftarrow \text{SEAT}(R_{ij}^c, \gamma_c^i, 0, 0)$  // sample child
2  if  $\pi = 0$ 
3    then  $i' \leftarrow \text{SEAT}(R_i^a, \gamma_a^i, 0, M)$  // sample adapter
4    if  $i' \neq i$ 
5      then  $\pi \leftarrow \text{SAMPLEIDENTIFICATION}(i', j, \mathbf{R}, \theta, H, M)$ 
6      else  $\pi \leftarrow \text{SEAT}(R^r, \gamma_r, d_r, H)$  // sample root
7  return  $\pi$ 

```

The nested, recursive *Chinese restaurant* construction serves to simulate further experiments, i.e., to sample more peptides given an already observed series π of peptides and will be useful in the following section to derive a likelihood function for parameter estimation.

2.4. Empirical Bayes parameter estimate

Parameters of the hierarchical process model from section 2.2 can be estimated from a series π of identifications by empirical Bayes inference, i.e., by choosing the parameters to maximize a likelihood function $\mathcal{L}_{\hat{\mathbf{R}}}$

$$\hat{\theta} := \arg \max_{\theta} \mathcal{L}_{\hat{\mathbf{R}}}(\theta) \quad (6)$$

In the following, we specify $\mathcal{L}_{\hat{\mathbf{R}}}$. Sampling a series π of identifications reduces to iterate various *Chinese restaurant* constructions according to the probabilities in (2). A likelihood function $\mathcal{L}_{\mathcal{R}}(\theta)$ for a set of seating arrangements \mathbf{R} , or the corresponding series π of identifications is defined as

$$\mathcal{L}_{\mathcal{R}}(\theta) = \mathcal{L}_{\text{cr}}(R^r, \gamma_r, d_r) \cdot \prod_{i=1}^I \mathcal{L}_{\text{cr}}(R_i^a, \gamma_a^i) \cdot \prod_{j=1}^J \mathcal{L}_{\text{cr}}(R_{ij}^c, \gamma_c^i) \quad (7)$$

where $\mathcal{L}_{\text{cr}}(R, \gamma, d)/\mathcal{L}_{\text{cr}}(R, \gamma)$ corresponds to the partial likelihood of achieving a seating arrangement R in a single restaurant representation of a Pitman-Yor/Dirichlet process sample with parameters $\gamma, d/\gamma$. Note that prior weights α_a^i of the adapter processes are appropriately incorporated into R_i^a and therefore not explicitly listed. The partial likelihood assumes the form

$$\mathcal{L}_{\text{cr}}(R, \gamma, d) = \frac{\prod_{k=1}^K (\gamma + kd) \cdot \prod_{n=1}^{n_k} (n - d)}{\prod_{n=1}^N (n + \gamma)} \quad (8)$$

with $N = \sum_{k=1}^K n_k$ and K corresponding to the number of populated tables.

We do observe the series π of identifications, although we only have incomplete knowledge about \mathbf{R} . We observe the seating arrangements R_{ij}^c of the child processes.

$$n_{ijk}^c = |\pi_t^i : (\pi_{t-1}^i = j) \wedge (\pi_t^i = k)| \quad (9)$$

where the $\pi_t^i \in \pi^i$ denote identifications observed exclusively in fraction i . R^r and the R_i^a cannot be directly observed. We present a sparse estimate for \mathbf{R} that is consistent with π and complies with a minimal number of seating events in the root restaurant representation R^r of the root distribution G . Consider the representation matrix \mathbf{M} with entries m_{ik} equaling one if a peptide k has been observed in fraction i or zero otherwise. Each peptide discovery k has to be represented by some fraction f_k . We further want to choose the number of representing fractions to be as small as possible. This problem is more commonly known as the NP-hard set cover problem (Karp, 1972). We compute the f_k with the greedy heuristic, choosing at each step the fraction which covers the largest number of remaining different peptides. Every time the peptide k is discovered, i.e., sampled for the first time in a child process, we choose the corresponding adapter process to trigger a sampling event in f_k . Accordingly, the hidden seating arrangements of the adapter (n_{ij}^a) and root restaurant representations (n_k^r) are estimated as follows:

$$\begin{aligned} n_{ij}^a &= |i, j, k : (f_k = i') \wedge (\exists t : (\pi_{t-1}^i = j) \wedge (\pi_t^i = k))| \\ n_k^r &= |i, j, k : (f_k = i) \wedge (\exists t : (\pi_{t-1}^i = j) \wedge (\pi_t^i = k))| \end{aligned} \quad (10)$$

We finally determine the parameters $\hat{\theta}$ by optimizing $\mathcal{L}_{\hat{\mathbf{R}}}$ with a quasi-Newton method (R Development Core Team, 2005). In summary, this protocol achieves an empirical Bayes parameter estimate from an observed series π of identifications.

2.5. Proteome coverage prediction with false identifications

At this point, it is possible to specify how to predict the number of new peptide discoveries for future experiments from a series π of already observed identifications. In a first step, parameters and hidden variables of the hierarchical process model (2.2) are estimated as described in the preceding section,

Section 2.4. Second, m peptide series $(\pi_{new,i})_{i=1}^m$ are sampled by means of the nested *Chinese restaurant* construction (2.3). For each $\pi_{new,i}$ the number of new discoveries is counted and the expected proteome coverage is estimated as the mean of discovery counts across all $\pi_{new,i}$.

In practice, the series π of observed peptides corresponds to a series of peptide-spectrum matches that have been inferred computationally. Obviously peptide-spectrum matches are not perfect. Fortunately, the fraction of false positive peptide-spectrum matches is typically known (Keller et al., 2002; Elias and Gygi, 2007). Furthermore, it has been observed that false positive peptide-spectrum matches distribute in a uniform-like manner across the protein database (Claassen et al., 2009; Reiter et al., 2009). To account for false positive peptide-spectrum matches, we adaptively estimate parameters and sample novel peptide identifications as described in Claassen et al., (2009).

3. RESULTS

We present results that demonstrate the proteome coverage prediction performance of our hierarchical process model. To this end, we studied a large multidimensional fractionation experiment of a *L. interrogans* sample. We compared our approach to a recent one designed for (one-dimensional) LC-MS/MS experiments (Claassen et al., 2009) and to ad hoc extrapolation methods. We further extrapolated proteome coverage for the *L. interrogans* sample to make statements about maximal coverage.

This study is based on an integrated dataset acquired from multidimensional fractionation experiments for the bacterium *L. interrogans*. After protein extraction and tryptic digestion, the resulting peptide mixture was fractionated according to the isoelectric point of the peptides by off-gel electrophoresis and each of the 24 fractions analyzed by LC-MS/MS coupled to a FT-LTQ high mass accuracy instrument. Target-decoy database search with Sequest/PeptideProphet (Keller et al., 2002) resulted in 59918 peptide-spectrum matches at a false discovery rate of 1% (Schmidt *et al.*, manuscript in preparation).

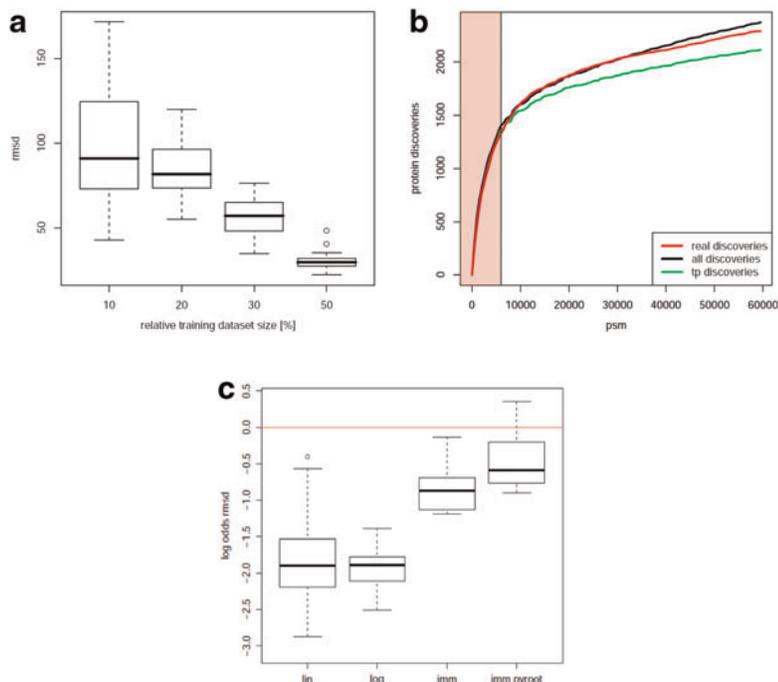
We assessed proteome coverage prediction performance in a cross validation scenario. Briefly, we generated various training datasets of decreasing size by subsampling the complete set of peptide-spectrum matches. We performed proteome coverage prediction for each training dataset and assessed accuracy by comparing to the real proteome coverage progression of the complete dataset. Precisely, we generated 20 training datasets by 20 times sampling 10% of all peptide-spectrum matches in the dataset while preserving their fraction association. We repeated this procedure by also sampling 20%, 30%, or 50% of all peptide-spectrum matches, finally obtaining 80 training datasets of varying size.

We assessed the prediction accuracy of the hierarchical process model (Fig. 2a). Prediction accuracy is measured as root mean square deviation of predicted and actually observed progression of proteome coverage. Proteome coverage corresponds to number of protein discoveries. Prediction accuracy is reasonable already for the smallest training dataset sizes, i.e., 10% of the complete *L. interrogans* dataset. Figure 2b depicts an example prediction for the set of smallest training datasets. As expected, prediction accuracy improves further for training datasets of larger size. Similar results are obtained for prediction of proteome coverage in terms of peptide discoveries (data not shown). These results demonstrate that our approach is able to reliably predict proteome coverage already from a small amount of data.

We compared the hierarchical process model to other methods, i.e., two simple general purpose extrapolation methods and a method designed for proteome coverage prediction of non-integrated datasets. We first considered an extrapolation scheme that linearly extrapolated proteome coverage progression of the last LC-MS/MS experiment of a training series. Second, we considered the extrapolation of a logarithmic regression ($y = a \log x + b$). Lastly, the hierarchical process model was benchmarked against an infinite Markov model based on Dirichlet process priors (Claassen et al., 2009) and a variant with a Pitman-Yor root process G (see Section 2.2). We assessed prediction performance on the 80 training series as described above and observed that the hierarchical process model clearly outperforms the other methods (Fig. 2c). These results indicate that proteome coverage prediction for integrated datasets is a non-trivial task that is not solved satisfactorily by ad hoc extrapolation methods and is different from the related task of proteome coverage prediction for non-integrated datasets.

We estimated saturation proteome coverage for *L. interrogans* given the experimental workflow described above. Therefore, we performed proteome coverage prediction for in silico repetition of all experiments. Proteome coverage in terms of peptide discoveries appears to steadily increase (Fig. 3a). Proteome coverage in terms of protein discoveries also seems to increase (Fig. 3b). This observation is

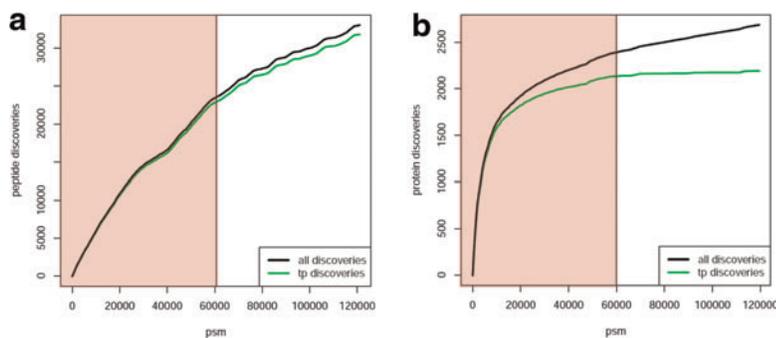
FIG. 2. Proteome coverage prediction performance by cross validation. Training datasets generated by subsampling the complete set of peptide-spectrum matches. Test of prediction performance on complete dataset. **(a)** Hierarchical process model accuracy in terms of root mean square deviation (rmsd) from the true progression of proteome coverage. Columns correspond to relative training dataset size compared to the complete *L. interrogans*. **(b)** Example trajectory for prediction from dataset instance with 10% relative size. Plot shows trajectory of observed (real), predicted true positive (tp) and including false positive protein discoveries (all). **(c)** Performance comparison of hierarchical process model with infinite Markov model (imm), infinite Markov model with Pitman-Yor root process (imm pyroot), extrapolation of logarithmic regression (log) and linear extrapolation of last experiment (lin). Box plot of log odds of rmsd ($\log(\text{rmsd}_{\text{ref}}/\text{rmsd}_{\text{comp}})$) for reference and compared method (lin, log, imm, imm pyroot). Median log odds for comparison with other methods are significantly lower than zero, indicating weaker performance than our approach. The hierarchical process model is capable to reliably predict proteome coverage from a small amount of identifications and clearly outperforms other applicable methods.



however only true for all protein discoveries including the false positive ones. Since our approach separately accounts for the contribution of false and true positive protein discoveries (see Section 2.5), we could exclusively monitor the progression of true protein discoveries. The number of true positive protein discoveries does not change significantly. Considering the rate of new true positive discoveries, we effectively have reached saturation coverage for *L. interrogans*.

We studied the proteome coverage progression for the *L. interrogans* dataset for protocols that select peptide-spectrum matches according to more stringent score thresholds. The last experiment demonstrated how the discrepancy between the progression of all and exclusively true protein discoveries grows with the number of acquired constant quality peptide-spectrum matches (Fig. 3b; peptide-spectrum match FDR 1%). More stringent peptide-spectrum match selection is supposed to translate into a smaller fraction of false

FIG. 3. Proteome coverage prediction beyond the *L. interrogans* dataset. Pink area denotes the extent of the dataset in terms of acquired peptide-spectrum matches (psm). Trajectories correspond to predicted true positive (tp) and including false positive discoveries (all). **(a)** Progression of peptide discoveries. **(b)** Progression of protein discoveries. Protein discovery rate stagnates compared to the steadily increasing number of peptide discoveries. The *L. interrogans* dataset achieves saturation coverage at the level of protein discoveries.



positive protein discoveries. This benefit comes at the cost of ignoring evidence for true discoveries (false negatives). We quantitatively studied this trade-off by comparing the expected coverage progression from the last experiment with the progression obtained for two other protocols that exclusively considered peptide-spectrum matches with FDR 0.5% (Fig. 4a) or 0.05% (Fig. 4b), respectively. The expected saturation coverage of the true protein discoveries remains virtually unchanged for any of the protocols. In contrast, the accumulation of false positive discoveries is reduced to insignificant levels for the most stringent selection protocol. These findings suggest that resorting to the most stringent selection protocol is expected to effectively achieve optimal proteome coverage at no significant accumulation of false protein discoveries. This suggestion independently confirms similar results by other studies on other datasets acquired for other organisms and by means of different experimental workflows (Reiter et al., 2009; Claassen et al., 2010).

4. CONCLUSION

We propose here a method to predict proteome coverage for multidimensional fractionation experiments. This achievement is an important enabling step for experimentalists since multidimensional fractionation experiments so far have the largest potential to comprehensively characterize a proteome. We present a novel hierarchical process to characterize distributions arising in the course of these experiments. This approach conceptually extends methods exclusively suited for single fraction experiments (Claassen et al., 2009), by introducing self-referential base measures that accommodate similarities among different experiment fractions. Our approach is generic since it operates on the level of peptide or protein distributions and, therefore, it conceptually accommodates any kind of heterogeneous set of fractions being analyzed by LC-MS/MS. Fractions do not necessarily have to originate from a single fractionation experiment. The considered fractions might also be derived from different tissues or cell cultures as long as their analysis is based on the same sequence database. Although we explicitly describe an approach that accounts for two fractionation steps, it is conceptually straightforward to extend it from a two level to a higher level hierarchy. However, the corresponding experimental setups are rarely encountered in practice. This study demonstrates that our model reliably predicts proteome coverage of future experiments from a small amount of already performed experiments and clearly outperforms other methods.

Besides providing predictions at the level of peptide discoveries, our approach yields reliable predictions of proteome coverage in terms of protein discoveries. Specifically, our approach requires the set of considered fragment ion spectra to be unambiguously assigned to a protein identity to estimate future proteome coverage. This requirement is usually met, since possible ambiguities introduced by peptide-spectrum matches whose sequence maps to several protein identities are typically resolved by protein inference engines, e.g., by reporting a minimal consistent set of protein identifications (Nesvizhskii et al., 2003). It will be interesting to extend our approach to allow for ambiguity in the protein identity assignments.

There has been considerable discussion in the past about when to consider a proteome to be mapped out. Our approach to proteome coverage prediction enables us to detect saturation coverage for any kind of

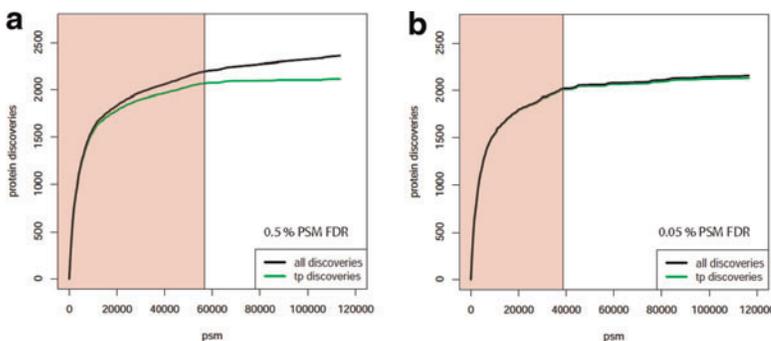


FIG. 4. Proteome coverage progression and prediction of sets of peptide spectrum matches selected at different levels of stringency: peptide spectrum match FDR at (a) 0.5%, and (b) 0.05%. More stringent selection of peptide spectrum matches comes along with a reduced number of considered peptide-spectrum matches (pink section). The more stringent the selection, the slower the observed

and predicted progression of total number of protein discoveries (black). The progression of true protein discoveries (green) saturates at virtually same level for all considered peptide-spectrum match selections.

shotgun proteomics dataset. In this study, the *L. interrogans* dataset reaches saturation coverage at the level of protein discoveries. Out of 3740 proteins reported in the sequence database, roughly 2000 proteins can be faithfully observed—not less but also not a lot more. This analysis is a remarkable result considering the manageable amount of experimentation (24 LC-MS/MS runs). It should be noted that this result is valid for the given experimental setup, such as type of protein extraction, enzymatic digestion, fractionation method, type of mass spectrometer. Despite the sensitive state-of-the-art approach reported here, it remains conceivable that other experimental approaches turn out to be able to explore other parts of the *L. interrogans* proteome. Their potential could though be evaluated with the hierarchical process model presented here. Therefore, the presented method is suited to assist method development since it objectively assesses the potential of a particular method to explore a proteome.

Characterizing more complex proteomes (e.g., human) necessitates a considerably larger amount of experimentation. In this context it will be promising to perform proteome coverage prediction for different experimental strategies at an early stage of the project to design future experiments such that maximal proteome coverage is achieved efficiently. Our approach enables for the first time to accommodate any multidimensional fractionation strategy to perform this task. Efficient study design will save costly experiments, contribute to the reliability of the final set of protein discoveries (Claassen et al., 2009; Reiter et al., 2009) and furthermore enhance subsequent directed/targeted proteomics studies (Schmidt et al., 2008; Lange et al., 2008).

ACKNOWLEDGMENTS

We thank Alexander Schmidt and Lukas Reiter for carefully reading the manuscript. We further thank Alexander Schmidt for kindly providing the *L. interrogans* data set. The project was supported in part by internal funds from ETH Zurich and by SystemsX.ch, the Swiss initiative for systems biology.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

- Antoniak, C.E. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Stat.* 2, 1152–1174.
- Beal, M.J., Ghahramani, Z., and Rasmussen, C.E. 2002. The infinite hidden Markov model. *Adv. Neural Inform. Process. Syst.* 1, 577–584.
- Blackwell, D., and MacQueen, J.B. 1973. Ferguson distributions via poly urn schemes. *Ann. Stat.* 1, 353–355.
- Brunner, E., Ahrens, C.H., Mohanty, S., et al. 2007. A high-quality catalog of the *Drosophila melanogaster* proteome. *Nat. Biotechnol.* 25, 576–583.
- Claassen, M., Aebersold, R., and Buhmann, J.M. 2009. Proteome coverage prediction with infinite Markov models. *Bioinformatics* 25, i154–i160.
- Claassen, M., Reiter, L., Hengartner, M.O., et al. 2010. Generic comparison of protein inference engine families. *Proc. RECOMB Satellite Comput. Proteomics* (in press).
- Domon, B., and Aebersold, R. 2006. Mass spectrometry and protein analysis. *Science* 312, 212–217.
- Elias, J.E., and Gygi, S.P. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 4, 207–214.
- Eriksson, J., and Fenyo, D. 2007. Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs. *Nat. Biotechnol.* 25, 651–655.
- Karp, R.M. 1972. Reducibility among combinatorial problems, 85–103. In Miller, R.E., and Thatcher, J.W., ed. *Complexity of Computer Computations*. Plenum Press, New York.
- Keller, A., Nesvizhskii, A.I., Kolker, E., et al. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 74, 5383–5392.
- Lange, V., Malmstrom, J.A., Didion, J., et al. 2008. Targeted quantitative analysis of *Streptococcus pyogenes* virulence factors by multiple reaction monitoring. *Mol. Cell Proteomics* 7, 1489–1500.

- Nesvizhskii, A.I., Keller, A., Kolker, E., et al. 2003. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 75, 4646–4658.
- Nesvizhskii, A.I., Vitek, O., and Aebersold, R. 2007. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* 4, 787–797.
- Pitman, J. 2002. Combinatorial stochastic processes. [Technical Report 621]. Department of Statistics, University of California, Berkeley.
- Pitman J., and Yor, M. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probabil.* 25, 855–900.
- R Development Core Team. 2005. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
- Reiter, L., Claassen, M., Schrimpf, S.P., et al. 2009. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell Proteomics* 8, 2405–2417.
- Schmidt, A., Gehlenborg, N., Bodenmiller, B., et al. 2008. An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. *Mol. Cell Proteomics* 7, 2138–2150.
- Teh, Y.W. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. *Proc. 21st Int. Conf. Comput. Linguistics* 985–992.
- Teh, Y.W., Jordan, M.I., Beal, M.J., et al. 2006. Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.* 101, 1566–1581.

Address correspondence to:

Dr. Manfred Claassen
Universitaetstrasse 6, CAB F 61.1
8092 Zurich, Switzerland

E-mail: manfred.claassen@inf.ethz.ch

