Challenges and opportunities for Bayesian statistics in proteomics

Oliver M. Crook * [1], Chun-wa Chung[2], and Charlotte M. Deane[1]

[1] *Department of Statistics, University of Oxford, Oxford, UK*
[2] *Structural and Biophysical Sciences, GlaxoSmithKline R&D, Stevenage, UK*

September 13, 2021

Abstract

Proteomics is a data-rich science with complex experimental designs and an intricate measurement process. To obtain insights from large datasets, statistical methodology and machine learning is routinely applied. For a quantity of interest, many of these approaches only produce a point estimate, such as a mean, leaving little room for bespoke interpretations. In contrast, Bayesian statistics quantifies uncertainty using probability distributions. These probability distributions allow scientist to ask complex questions of their proteomics data which would otherwise be challenging using alternative approaches. Bayesian statistics also offers a modular framework for specifying complex hierarchies of parameter dependencies. This allows us to use statistical methodology which equals, rather than neglects, the sophistication of experimental design and instrumentation present in proteomics. Here, we review Bayesian methods applied to proteomics and argue for a broader uptake, whilst also highlighting the challenges posed by adopting a new statistical framework. To illustrate our review, we present a walk-through of the development of a Bayesian model for dynamic organic orthogonal phase-separation (OOPS) data.

1 Introduction

Decision making spans the entire research process. Ultimately, it is a choice to believe an explanation for a phenomena given the current evidence. For some theories, the evidence is overwhelming: careful mechanistic experiments and verifiable model predictions have never contradicted that theory. This scenario is, however, rare. In practice, we make decisions under uncertainty and the evidence is not clear-cut. Bayesian statistics allows us to make

*oliver.crook@stats.ox.ac.uk

inferences from that evidence to enable decision making in those cases. In contrast to *frequentist* methods, Bayesian inference allows us to use probability to model degrees of belief, rather than just frequencies. Consequently, models that are consistent with the available evidence are more probable and incompatible models are less probable. By using probability theory in this manner, there is a recipe for taking *prior beliefs* (i.e. information encoded by domain expertise) and updating them to *posterior beliefs* using observed data. As a result, this posterior probability distribution quantifies the models compatible with domain expertise and our experimental data. This recipe is known more formally as *Bayes' theorem*.

Mass-spectrometry-based proteomics is a complex scientific field. The techniques versatility allows us to explore differential abundance, protein turnover, interactions, thermal stability, structure, spatial information and more. In each case, data are manipulated, thresholded and filtered so that a statistical test or machine learning algorithm can be applied. The results are then frequently concluded with a single value, which we have granted the role of arbiter of truth. These decisions are often made without consideration of what we might be happening at each step. Bayesian statistics could propagate or quantify the uncertainty in these steps, replace implicit or ad-hoc approaches with explicit models and summarise the output with a probability distribution consistent with our data. This paradigm progression not only provides us an ability to ask new questions of our data but a consistent way to perform inference and criticize our models.

Bayesian statistics offers proteomics considerable possibilities; despite that, it has not been readily adopted in the community. This may stem from a lack of familiarity, a lack of awareness of available tools, complex language, impenetrable literature, inability to communicate results from an analysis and computational difficulties. Here, we review the contribution Bayesian statistics has already made to proteomics, clarify the Bayesian workflow and how it can be applied to proteomics, highlight a number of modelling strategies and outline current challenges for the community. Throughout, we illustrate our analysis with examples from the proteomics literature, focusing on building a model for dynamic organic orthogonal phase separation data.

2 Main

2.1 Bayes, in brief

Before we review the contributions Bayesian statistics has already made to proteomics, we introduce the fundamental technical background and notation. We use $P(E)$ to denote the probability of the event E. E can be anything from "it rains tomorrow" to "my parameter falls between the values a and b". We let D be notation for the observed data, for example from a shotgun proteomics experiment. Let x be a data point from D, such as a measurement for a particular protein. We assume that x arises from some probability distribution p and we write $x \sim p(x|\theta)$, for example this could be a log normal distribution. Let α be

hyperparameters of the parameter distribution, such that θ themselves are drawn from a probability distribution $\theta \sim p(\theta|\alpha)$. For example, the mean of the log normal distribution could be drawn from a normal distribution.

The *prior distribution* captures our domain expertise and is the distribution of the parameters before any data is observed: $p(\theta|\alpha)$. The *prior* could capture, say, that abundance values are positive and are unlikely to exceed the number of grains of sand on Earth. The sampling distribution is the distribution of the data given the parameters $p(D|\theta)$, we can write this as a function $L(\theta|D)$ called the *likelihood*. If we average (or *marginalise*) the distribution of the data over the parameters, we obtain the so-called marginal likelihood:

$$p(D|\alpha) = \int L(\theta|D)p(\theta|\alpha)\,\mathrm{d}\theta. \tag{1}$$

The posterior distribution of the parameters is determined by Bayes' theorem, as the following:

$$p(\theta|D,\alpha) = \frac{p(D|\theta,\alpha)p(\theta|\alpha)}{p(D|\alpha)}. \tag{2}$$

Bayes' theorem tells us the mathematical way to update beliefs in light of evidence: simply multiply our prior and likelihood and renormalise by the marginal likelihood. Bayes' theorem implies a self-consistency property: the posterior averaged over the data returns the prior:

$$p(\theta) = \int\int p(\theta|\tilde{y})p(\tilde{y}|\tilde{\theta})p(\tilde{\theta})\,\mathrm{d}\tilde{y}\mathrm{d}\tilde{\theta}, \tag{3}$$

where $\tilde{y} \sim p(D|\tilde{\theta})$. When Bayesian's predict, instead of simply taking a single parameter value forward, they make predictions by averaging:

$$p(\tilde{x}|D,\alpha) = \int p(\tilde{x}|\theta)p(\theta|X,\alpha)\,\mathrm{d}\theta. \tag{4}$$

In summary, Bayesian statistics provides us with a distribution of plausible parameter values from the *posterior* and a distribution of hypothetical predicted values from the *posterior predictive distribution*. We can then ask bespoke question of these probability distributions; for example, $P(\theta > 2|D,\alpha) = \int_2^\infty p(\theta|D\,\alpha)\,\mathrm{d}\theta$ is the probability that a parameter is greater than 2. For proteomics, these could be the probability that a fold-change exceeded a certain value or a the probability that a spectra belongs to a particular peptide. One possible interpretation of the prior is as a penalty that regularises the parameters, many applications exploit this and so can apply bespoke regularisation to their model.

2.2 Bayesian contributions to proteomics

2.2.1 Bottom-up proteomics and differential abundance

Here, we highlight Bayesian approaches already applied to proteomics. We focus on proteomics data generated via mass-spectrometry but refer to contributions for the reverse-phase-protein-array (RPPA) literature (Crook *et al.*, 2019; Ni *et al.*, 2019; Maity *et al.*,

2020) and 2D gel electrophoresis (Morris *et al.*, 2011). A number of approaches have been aimed at quantification and differential abundance analysis of bottom-up proteomics data (Phillips *et al.*, 2021; The and Käll, 2021, 2019; Santra and Delatola, 2016; Peshkin *et al.*, 2019; Millikin *et al.*, 2020; Serang *et al.*, 2013; O'Brien *et al.*, 2018; Serang *et al.*, 2012; Carvalho *et al.*, 2011). Carvalho *et al.* (2011) appreciate that Bayesian statistics can be used to improve analysis of spectral counting data using a Poisson likelihood and calculating the probability of detecting a protein in a particular sample. However, they do not exploit the full Bayesian toolkit by working with probability distributions and simply end-up interpreting their probabilities as p-values. Thus, it is not clear whether their approach is any better than simply using a likelihood-based approach. A number of approaches (The and Käll, 2021, 2019; Peshkin *et al.*, 2019; Serang *et al.*, 2012) argue for propagating and quantifying the uncertainty in the analysis rather than simply the underlying quantitation values, either through relaxing the parsimony assumption (Serang *et al.*, 2012), including ion statistics via a two-level Beta-Binomial model (Peshkin *et al.*, 2019), or jointly modelling identification and quantitation statistics (The and Käll, 2019, 2021). Millikin *et al.* (2020) argue to use an analogue of the t-test and exploit the posterior distribution by using an interval thresholding approach. O'Brien *et al.* (2018) note the inherent compositional nature of labelled proteomics approaches and include a modelling parameter to model ratio compression allowing better estimation of the true fold changes. Importantly this parameter is shared across all proteins, which allowed them to estimate the parameter accurately even for proteins with few observed peptides. Jow *et al.* (2014) also model isobaric laballed mass-spectrometry data but do not model ratio compression. Meanwhile for label-free experiments, O'Brien *et al.* (2018) explicitly model missingness showing that jointly modelling missingness and abundance leads to improved performance. All of these approaches demonstrate some benefit over previously applied approaches suggests that combining these method would provide further improvements. It also suggests translating these methods to other proteomics techniques would be a fruitful endeavour. However, many of the methods differ in a number of key modelling choices and it is not always clear how these choices were made.

2.2.2 Protein and peptide identification

One of the fundamental problems in mass-spectrometry based proteomics is identifying a peptide from a spectra. A spectra can be very noisy and $b-, y-$ ions can be missing which results in a complex observation process. Furthermore, we have prior knowledge of observing particular amino acid sequences and knowledge of the cleavage process. This appears an ideal scenario for application of Bayesian methods. Indeed, a number of approaches have been applied (Chen *et al.*, 2005; Halloran *et al.*, 2016; Lewis *et al.*, 2018; Claassen *et al.*, 2009). Chen *et al.* (2005) use a fairly simple framework to calculate peptide identification probabilities based on peptide coordinates. Halloran *et al.* (2016) employ a dynamic Bayesian network in a method called DRIP, allowing for insertions and deletions to the spectra. By modelling possible alignments between theoretical and observed spectra they can calculate the most

probable peptide match. The authors find their approach improves over available methods particularly for low resolution MS2 data. Lewis *et al.* (2018) take a different approach to the same problem, incorporating a scoring function into a likelihood model. Their model also allows deletions directly via indicator functions. Insertions are characterised by excessive deviations from the spectra of the candidate peptide, where excessive is characterised probabilistically via laplace noise. The authors also include prior information about possible cleavage pairs, as well as prior information about the probability of observing a particular peptide sequence in the dataset. Finally, in constrast to Halloran *et al.* (2016) they make full use of Bayesian methods and provide a posterior distribution over possible peptides and parameters. This could allow multiple peptides to be associated to a spectrum with differing certainty which could be used in downstream analysis. Claassen *et al.* (2009) tackle a slightly different problem and use a non-parametric Bayesian model to predict the coverage in sequential LC-MS/MS experiments but suggest their approach could also be adapted to database searching and de novo sequencing. Again, these approaches all have benefits over previously applied methodology. The clearest is the inclusion of more information and the ability to provide a flexible, and well rationalised, model to the underlying data. The ability to exploit uncertainty captured by the posterior distribution for downstream analysis is far more insightful that simply point estimates from a Bayesian analysis. However, these approaches have not been adopted by the community. This could be because the methods are difficult to apply, the benefits are not compelling or computation is excessive.

2.2.3 Proteoforms and post-translation modifications

A number of approaches are interested in applications to proteoform analysis (splice isoforms) or post-translational modifications (Chung *et al.*, 2013; Webb-Robertson *et al.*, 2014; Lim *et al.*, 2017; Shteynberg *et al.*, 2019; Mallikarjun *et al.*, 2020). Chung *et al.* (2013) employ a non-parametric mixture model to jointly model the modification mass for each PTM group and the true (unobserved) location of the modified amino acid. Their approach outperforms other approaches, is fully automated and provides modification confidence scores. However, the approach does not model the underlying spectrum, which could result in unnecessary false positives. Webb-Robertson *et al.* (2014) tackle the proteoform problem by deconvolving peptides into signatures even if they are associated with the same protein. However, they only use a Bayesian point estimate rather than exploiting the full posterior distribution. Lim *et al.* (2017) use a Bayesian model to estimate the phosphorylation stoichiometry using Bayesian statistics. By incorporating a physically plausible model they remove problems with previous models that could allow negative stoichiometry. Their joint model allows them to borrow power across replicates and they report downstream uncertainty. Shteynberg *et al.* (2019) use a Bayesian mixture model to compute compute probabilities for modification sites. This allows them to combine precomputed scores in a rational way but again they do not examine the full posterior distributions. Mallikarjun *et al.* (2020) employ a Bayesian linear regression modelling strategy to analyse differential PTM data, suggesting their approach

outperformed other methods and could allow uncertainty in missing values. The main benefit here appears to be the regularisation of the parameters using priors rather than specifically the uncertainty quantification in the analysis.

2.2.4 Biomarkers and clinical proteomics

Protein biomarkers, molecular indicators of aberrant processes or disease, and clinical proteomics are a key component of proteomics research. Hernández *et al.* (2015) provide a review of Bayesian method development in biomarker development and we refer to them for additional details. Morris *et al.* (2006, 2008) develop Bayesian wavelet-based functional mixed models for mass-spectrometry-based proteomics data. Their advanced framework, allows the simultaneous use of nonparametric fixed and random effects, which facilitates adjustment for clinical and experimental covariates that could affect the intensity and location of a spectra. Working with posterior distributions they are able to compute important quantities such as the probability of intensity changes for fixed fold levels and are able to control a Bayesian false discovery rate. Liao *et al.* (2014) combine the above framework with image analysis methods to enable biomarker discovery from LC-MS data. Hwang *et al.* (2008) develop a pipeline, MS-BID, for biomarker analysis which uses a Bayesian anova. Harris *et al.* (2009) apply a Bayesian hierarchical linear probit regression model to determine discriminative biomarkers from mass spectrometry data. They find their approach improves over a K-nearest neighbour method. Furthermore, by using posterior probabilities they are able to determine which samples will the most promising for prognostics. Kuschner *et al.* (2010) propose a Bayesian network to perform feature selection from mass-spectrometry data. The selected features then provided excellent predictive power. Though again this approach still uses a Bayesian point estimate and could have obtained more information by computing full posterior distributions. Deng *et al.* (2007) develop a Bayesian network which allows them to integrate mass-spectrometry and microarray data, allowing them to borrow power between mRNA and protein levels. Here Bayesian statistics is sufficiently flexible to incorporate different modalities and weigh up the uncertainty between different datasets. More recently Liu *et al.* (2020) developed Bayesian Function-on-Scalar quantile regression for mass-spectrometry data. This approach notes that biomarker difference may not be apparent at mean regression but rather at a particular quantile (such as the 0.95 quantile). This method simultaneously accounts for the functional nature of MALDI-TOF data, incorporates prior knowledge for adaptive regularization and a basis representation which allows borrowing of power. They find their method identifies biomarkers overlooked by mean regression. Bayesian methods for biomarker and clinical proteomics are more developed than other examined proteomics sub-fields with several exemplary methods that make full use of the flexibility of Bayesian modelling and the rich output of the posterior distribution.

2.2.5 Chromatography

To facilitate identification in mass-spectrometry a liquid-chromatography step is usually applied. The time at which a peptide elutes from the liquid chromatography, called the retention time, can be used as additional information to help identify peptides. However, there is uncertainty in this retention time and they can vary from one run to another. Chen *et al.* (2019) develop a Bayesian model called DART-ID, which models a latent (unobserved) global retention time alignment. This alignment allows them to combine the outputted posterior error probability of MaxQuant with the inferred RT density in each experiment. Hence, by using this result they can update their confidences and improve coverage in experiments by 50%. Whilst their approach appears powerful, they only use a point estimate and obtain uncertainty through bootstrapping. Though the author reference computational challenges it would have been useful to examine the posterior distribution so we could be explicit about the computational trade-offs. Maboudi Afkham *et al.* (2017) are interested in the uncertainty in peptide retention time methods. Using a Gaussian process regression method they were able to accurately predict retention times and obtain uncertainty estimates. They then use the posterior distribution from the regression analysis as a variable retention time window to identify potentially incorrect peptides. This improves over fixed windowing strategies. One strategy they overlooked, in a similar vain to DART-ID, would be to update the identification probabilities based the deviation probability from the predicted retention time. This approach naturally fits within a Bayesian framework.

2.2.6 Intact, top-down and structural proteomics

Proteoform analysis is one of the key challenges in top-down proteomics. LeDuc *et al.* (2014) introduce a C-score, not be confused with a C-statistic, to facilitate automated identification and characeterisation of proteoforms from top-down proteomics data. Ultimately, their approach allows them to rank probable proteoforms having observed their data. Performing an analysis in a Bayesian framework allows them to specify a generative model, provide expert prior information and carefully model the underlying noise distribution. Their proposed C-score is essentially a transformed posterior error probability. However, despite their Bayesian framework, they opt for a point estimate of their model, which could have been greatly enhanced by examining the full posterior of their model. Marty *et al.* (2015) proposed a Bayesian deconvolution algorithm for Ion Mobility spectra, and extended in Kostelic *et al.* (2021). Their approach allows the convolution of the charge distribution with the peak shape to obtain a flexible deconvolution approach. The extent of their applications is extensive, demonstrating a clear benefit of their method. However, their approach also uses a point estimate from their analysis. Indeed, the idea of a posterior distribution is not mentioned at all within the paper. Hence, apart from the use of prior information, it is not clear what particular benefit a Bayesian analysis is for their approach.

2.2.7 Functional proteomics

References

Carvalho, P. C. et al. (2011). Analyzing marginal cases in differential shotgun proteomics. *Bioinformatics*, **27**(2), 275–276.

Chen, A. T. et al. (2019). Dart-id increases single-cell proteome coverage. *PLoS computational biology*, **15**(7), e1007082.

Chen, S. S. et al. (2005). Improving mass and liquid chromatography based identification of proteins using bayesian scoring. *Journal of proteome research*, **4**(6), 2174–2184.

Chung, C. et al. (2013). Non-parametric bayesian approach to post-translational modification refinement of predictions from tandem mass spectrometry. *Bioinformatics*, **29**(7), 821–829.

Claassen, M. et al. (2009). Proteome coverage prediction with infinite markov models. *Bioinformatics*, **25**(12), i154–i160.

Crook, O. M. et al. (2019). Fast approximate inference for variable selection in dirichlet process mixtures, with an application to pan-cancer proteomics. *Statistical applications in genetics and molecular biology*, **18**(6).

Deng, X. et al. (2007). Cross-platform analysis of cancer biomarkers: a bayesian network approach to incorporating mass spectrometry and microarray data. *Cancer informatics*, **3**, 117693510700300001.

Halloran, J. T. et al. (2016). Dynamic bayesian network for accurate detection of peptides from tandem mass spectra. *Journal of proteome research*, **15**(8), 2749–2759.

Harris, K. et al. (2009). Definition of valid proteomic biomarkers: a bayesian solution. pages 137–149.

Hernández, B. et al. (2015). Bayesian methods for proteomic biomarker development. *EuPA Open Proteomics*, **9**, 54–64.

Hwang, D. et al. (2008). Ms-bid: a java package for label-free lc-ms-based comparative proteomic analysis. *Bioinformatics*, **24**(22), 2641–2642.

Jow, H. et al. (2014). Bayesian identification of protein differential expression in multi-group isobaric labelled mass spectrometry data. *Statistical applications in genetics and molecular biology*, **13**(5), 531–551.

Kostelic, M. et al. (2021). Unideccd: Deconvolution of charge detection-mass spectrometry data.

Kuschner, K. W. et al. (2010). A bayesian network approach to feature selection in mass spectrometry data. *BMC bioinformatics*, **11**(1), 1–10.

LeDuc, R. D. et al. (2014). The c-score: a bayesian framework to sharply improve proteoform scoring in high-throughput top down proteomics. *Journal of proteome research*, **13**(7), 3231–3240.

Lewis, N. H. et al. (2018). Peptide refinement by using a stochastic search. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **67**(5), 1207–1236.

Liao, H. et al. (2014). A new paradigm for clinical biomarker discovery and screening with mass spectrometry through biomedical image analysis principles. pages 1332–1335.

Lim, M. Y. et al. (2017). Improved method for determining absolute phosphorylation stoichiometry using bayesian statistics and isobaric labeling. *Journal of proteome research*, **16**(11), 4217–4226.

Liu, Y. et al. (2020). Function-on-scalar quantile regression with application to mass spectrometry proteomics data. *The Annals of Applied Statistics*, **14**(2), 521–541.

Maboudi Afkham, H. et al. (2017). Uncertainty estimation of predictions of peptides' chromatographic retention times in shotgun proteomics. *Bioinformatics*, **33**(4), 508–513.

Maity, A. K. et al. (2020). Bayesian data integration and variable selection for pan-cancer survival prediction using protein expression data. *Biometrics*, **76**(1), 316–325.

Mallikarjun, V. et al. (2020). Bayesenproteomics: Bayesian elastic nets for quantification of peptidoforms in complex samples. *Journal of proteome research*, **19**(6), 2167–2184.

Marty, M. T. et al. (2015). Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. *Analytical chemistry*, **87**(8), 4370–4376.

Millikin, R. J. et al. (2020). A bayesian null interval hypothesis test controls false discovery rates and improves sensitivity in label-free quantitative proteomics. *Journal of proteome research*, **19**(5), 1975–1981.

Morris, J. S. et al. (2006). Analysis of mass spectrometry data using bayesian wavelet-based functional mixed models. bayesian inference for gene expression and proteomics.

Morris, J. S. et al. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics*, **64**(2), 479–489.

Morris, J. S. et al. (2011). Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *The annals of applied statistics*, **5**(2A), 894.

Ni, Y. et al. (2019). Bayesian hierarchical varying-sparsity regression models with application to cancer proteogenomics. *Journal of the American Statistical Association*, **114**(525), 48–60.

O'Brien, J. J. et al. (2018). Compositional proteomics: Effects of spatial constraints on protein quantification utilizing isobaric tags. *Journal of proteome research*, **17**(1), 590–599.

O'Brien, J. J. et al. (2018). The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *The annals of applied statistics*, **12**(4), 2075.

Peshkin, L. et al. (2019). Bayesian confidence intervals for multiplexed proteomics integrate ion-statistics with peptide quantification concordance. *Molecular & Cellular Proteomics*, **18**(10), 2108–2120.

Phillips, A. et al. (2021). Uncertainty aware protein-level quantification and differential expression analysis of proteomics data with seamass. *Statistical methods for proteomics*.

Santra, T. et al. (2016). A bayesian algorithm for detecting differentially expressed proteins and its application in breast cancer research. *Scientific reports*, **6**(1), 1–10.

Serang, O. et al. (2012). Recognizing uncertainty increases robustness and reproducibility of mass spectrometry-based protein inferences. *Journal of proteome research*, **11**(12), 5586–5591.

Serang, O. et al. (2013). Nonparametric bayesian evaluation of differential protein quantification. *Journal of proteome research*, **12**(10), 4556–4565.

Shteynberg, D. D. et al. (2019). Ptmprophet: fast and accurate mass modification localization for the trans-proteomic pipeline. *Journal of proteome research*, **18**(12), 4262–4272.

The, M. et al. (2019). Integrated identification and quantification error probabilities for shotgun proteomics. *Molecular & cellular Proteomics*, **18**(3), 561–570.

The, M. et al. (2021). Triqler for maxquant: Enhancing results from maxquant by bayesian error propagation and integration. *Journal of proteome research*, **20**(4), 2062–2068.

Webb-Robertson, B.-J. M. et al. (2014). Bayesian proteoform modeling improves protein quantification of global proteomic measurements. *Molecular & cellular proteomics*, **13**(12), 3639–3646.