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 6, 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 value 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 prediction by averaging:

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

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.

2.2 Bayesian contributions to proteomics

Here, we highlight Bayesian approaches already applied to proteomics. We focus on proteomics data generated via mass-spectrometry but refer to excellent contributions for the reverse-phase-protein-array (RPPA) literature.