EuPA Open Proteomics 9 (2015) 54–64

Contents lists available at ScienceDirect

EuPA Open Proteomics

journal homepage: www.elsevier.com/locate/euprot

Bayesian methods for proteomic biomarker development

Belinda Hernández[a,b,*], Stephen R Pennington[a,b,1], Andrew C Parnell[a,c,1]

[a] School of Mathematical Sciences (Statistics), University College Dublin, Belfield Campus, Dublin 4, Ireland
[b] School of Medicine and Medical Science, UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield Campus, Dublin 4, Ireland
[c] Insight: The National Centre for Data Analytics, University College Dublin, Belfield Campus, Dublin 4, Ireland

ARTICLE INFO

Article history:
Received 12 January 2015
Received in revised form 18 May 2015
Accepted 6 August 2015
Available online 10 August 2015

Keywords:
Bayesian statistics
R
proteomics biomarker discovery
LC–MS

ABSTRACT

The advent of liquid chromatography mass spectrometry has seen a dramatic increase in the amount of data derived from proteomic biomarker discovery. These experiments have seemingly identified many potential candidate biomarkers. Frustratingly, very few of these candidates have been evaluated and validated sufficiently such that that they have progressed to the stage of routine clinical use. It is becoming apparent that the statistical methods used to evaluate the performance of new candidate biomarkers are a major limitation in their development. Bayesian methods offer some advantages over traditional statistical and machine learning methods. In particular they can incorporate external information into current experiments so as to guide biomarker selection. Further, they can be more robust to over-fitting than other approaches, especially when the number of samples used for discovery is relatively small.

In this review we provide an introduction to Bayesian inference and demonstrate some of the advantages of using a Bayesian framework. We summarize how Bayesian methods have been used previously in proteomics and other areas of bioinformatics. Finally, we describe some popular and emerging Bayesian models from the statistical literature and provide a worked tutorial including code snippets to show how these methods may be applied for the evaluation of proteomic biomarkers.

Contents

1. Introduction . . . 55
2. What is Bayesian inference? . . . 55
3. Motivation for using Bayesian methods . . . 55
4. Bayesian models currently used in proteomics . . . 56
   4.1. Biomarker discovery . . . 56
   4.2. Other areas of proteomics . . . 56
5. Possible Bayesian applications for biomarker discovery . . . 56
   5.1. Bayesian Lasso . . . 57
   5.2. Other Priors for Variable Selection . . . 57
   5.3. Bayesian non-parametric models . . . 58
      5.3.1. Bayesian CART . . . 58
      5.3.2. Other Bayesian tree models . . . 58
      5.3.3. Bayesian additive regression trees (BART) . . . 58
   5.4. Worked example: implementation of Bayesian inference for biomarker evaluation in R . . . 59
      5.4.1. Package reglogit . . . 59
      5.4.2. JAGS . . . 60
      5.4.3. bartMachine . . . 61
6. Discussion . . . 62

* Corresponding author.
E-mail address: Belinda.Hernandez@ucdconnect.ie (B. Hernández).
[1] Joint contribution.

http://dx.doi.org/10.1016/j.euprot.2015.08.001

Acknowledgements 62
References 62

1. Introduction

Advances in proteomic technology, in particular the widespread use of liquid chromatography mass spectrometry (LC–MS), have meant that vast amounts of information regarding protein and peptide features can now be easily collected from bodily fluids and tissue, making them an ideal target to find biomarkers of disease. A mass spectrum sample can be represented as a series of peaks where the mass to charge ratio (m/z) is depicted on the x-axis and the molecule intensity on the y-axis. In statistical and bioinformatic analysis each m/z ratio is treated as a separate variable where its value is the intensity or abundance of the molecule at the given m/z ratio. Each peak generally corresponds to a protein fragment or peptide and so the objective of most biomarker discovery experiments is to find a subset of peptides that best discriminate between the outcome groups [1]. It is widely accepted that use of individual biomarkers are unlikely to sufficiently capture the complexity and possible heterogeneity of a given disease [2–4]. For this reason, most studies focus on finding a panel or signature of differentially expressed protein or peptide features that are both sensitive and specific enough to accurately predict a treatment or disease state.

It has now become clear that the issue of finding a sensitive and specific panel of biomarkers is much more complex than initially anticipated. The area of proteomic biomarker discovery was initially met with high hopes and great enthusiasm; however, this fervor has weaned in recent years due to the inability of many studies to validate candidate biomarkers that were initially thought to be highly discriminatory [5,6]. Because of this, few proteomic biomarkers have reached clinical utility despite much government and industry investment [7,8]. Many articles have reflected on the shortcomings of these earlier studies and have laid out guidelines to rectify the oversights of initial experiments [7,9,10].

Bayesian methods have been widely used in many areas of bioinformatics and proteomics mainly due to the fact that they lend themselves nicely to the challenge of analyzing complex, noisy and often incomplete data [11]. Their growing popularity over the last 20 years is mainly attributable to advances in computational power which make fitting Bayesian models much more attainable for large datasets [12]. This article reviews the literature on Bayesian methods in proteomics in general before focusing on how Bayesian methods can be used for the statistical analysis of mass spectra data for proteomic biomarker discovery and evaluation. The benefits of using Bayesian models compared to other traditional and machine learning methods are discussed. Reasons why Bayesian models might attain superior performance in the validation of separate cohorts are also identified. Furthermore we highlight methods used in other areas of research as well as other recent developments in Bayesian analysis which could prove to be useful in future applications of proteomic biomarker discovery experiments. Section 5.4 is a tutorial comprising a worked proteomic validation example, using candidate biomarkers for the prediction of cardiovascular disease, in which two of these Bayesian methods are tested and compared against each other. This section also includes code samples which may be used to run these models using freely available software. The tutorial data set is also provided in order for the reader to test the code provided and run the tutorial in their own time. Those not interested in following the tutorial may wish to skip Section 5.4 and proceed directly to Section 6.

2. What is Bayesian inference?

At the heart of all Bayesian methods is Bayes' theorem,

$$p(\theta|Y) \propto p(\theta) \times p(Y|\theta)$$

often expressed in words as:

posterior is proportional to prior times likelihood.

In the above equation Y is the experimental data and θ are the unknown parameters (e.g., peptide importance values). The posterior distribution $p(\theta|Y)$is the joint probability distribution of the unknown parameters given the observed data. Bayes' theorem states that the posterior distribution can be calculated from a combination of a probability distribution on the unknown parameters of interest $p(\theta)$ known as the prior distribution and a conditional probability distribution $p(Y|\theta)$of the data Y given the parameters θ, known as the likelihood.

Commonly, the prior distribution $p(\theta)$ represents the knowledge about the parameters of interest θ before any data is collected. Its shape represents the degree of certainty or knowledge about θ; for example a distribution with a sharp peak would express high confidence in our knowledge of θ whereas a flat or uninformative prior would express no prior knowledge about the parameters of interest. When data become available after an experiment has been conducted, the information about the data and the parameters of interest are combined through Bayes' theorem to produce $p(\theta|Y)$. The main aim of any Bayesian analysis is to identify a credible set of values that the parameters θ can take given the observed data Y [12,13], i.e., find the posterior distribution.

Bayes' formula is written with a proportionality constraint $(\propto)$ because the full equation involves calculating a difficult integral, known as the normalizing constant. This problem is neatly sidestepped by using fitting methods such as Markov Chain Monte Carlo (MCMC) which make inference about the posterior distribution by sampling from it rather than computing it explicitly.

3. Motivation for using Bayesian methods

One of the main advantages of Bayesian methods over non-Bayesian statistical and machine learning techniques is the ability to incorporate external information about the parameters through the prior distribution. In proteomic experiments in particular a great deal is already known about the parameters of interest before an experiment takes place which can be incorporated into the prior distribution. For example, if it was known that certain peptide features tend to have high technical variability and be less reproducible (as is often the case in MS analysis with low abundant features whose intensity is near the limit of detection of the mass spectrometer) a less informative prior could be used on these peptides as opposed to the higher abundance, more reproducible features.

One of the main reasons for the failure of many initial discovery studies to validate according to [14] is the failure to accurately model sources of experimental and biological variability. Many traditional pre-existing techniques have been used to analyze the data resulting from proteomic biomarker experiments such as support vector machines, random forests, Lasso regression and various other classification methods [15–17]. However, the one disadvantage common to all these methods is that they ignore the uncertainty introduced to the data and assume that the