

1 *Bayesian Methods for Proteomic*  
2 *Biomarker Discovery*

---

3

4

5

6 Belinda Hernández<sup>1,3</sup>, Stephen R Pennington\*<sup>3</sup>, Andrew C Parnell \*<sup>1,2</sup>

7

8 <sup>1</sup> School of Mathematical Sciences (Statistics), University College Dublin, Belfield Campus, Dublin 4,  
9 Ireland

10 <sup>2</sup>Insight: The National Centre for Data Analytics, University College Dublin, Belfield Campus, Dublin 4,  
11 Ireland

12 <sup>3</sup> School of Medicine and Medical Science, UCD Conway Institute of Biomolecular and Biomedical  
13 Research, University College Dublin, Belfield Campus, Dublin 4, Ireland

14 \*Joint contribution

15

16 **Corresponding Author**

17 Email: [Belinda.Hernandez@ucdconnect.ie](mailto:Belinda.Hernandez@ucdconnect.ie)

18

19

## 20 **Abstract**

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

30

31 In this review we provide an introduction to Bayesian inference and demonstrate some of the  
32 advantages of using a Bayesian framework. We summarise how Bayesian methods have been used  
33 previously in proteomics and other areas of bioinformatics. Finally, we describe some popular and  
34 emerging Bayesian models from the statistical literature and provide a worked tutorial including  
35 code snippets to show the reader how to use these methods for the discovery of proteomic  
36 biomarkers in their own experiments.

37

## 38 **1 Introduction**

39 Advances in proteomic technology, in particular the widespread use of liquid chromatography mass  
40 spectrometry (LC-MS), have meant that vast amounts of information regarding protein and peptide  
41 features can now be easily collected from bodily fluids and tissue, making them an ideal target to  
42 find biomarkers of disease. A mass spectrum sample can be represented as a series of peaks where  
43 the mass to charge ratio ( $m/z$ ) is depicted on the x-axis and the molecule intensity on the y-axis. In  
44 statistical and bioinformatic analysis each  $m/z$  ratio is treated as a separate variable where its value  
45 is the intensity or abundance of the molecule at the given  $m/z$  ratio. Each peak generally  
46 corresponds to a protein fragment or peptide and so the objective of most biomarker discovery  
47 experiments is to find a subset of peptides that best discriminate between the outcome groups [1].  
48 It is widely accepted that use of individual biomarkers are unlikely to sufficiently capture the  
49 complexity and possible heterogeneity of a given disease [2][3][4]. For this reason, most studies  
50 focus on finding a panel or signature of differentially expressed protein or peptide features that are  
51 both sensitive and specific enough to accurately predict a treatment or disease state.

52 It has now become clear that the issue of finding a sensitive and specific panel of biomarkers is much  
53 more complex than initially anticipated. The area of proteomic biomarker discovery was initially met  
54 with high hopes and great enthusiasm; however this fervor has waned in recent years due to the  
55 inability of many studies to validate candidate biomarkers that were initially thought to be highly  
56 discriminatory [5][6]. Because of this, few proteomic biomarkers have reached clinical utility despite  
57 much government and industry investment [7][8]. Many articles have reflected on the shortcomings  
58 of these earlier studies and have laid out guidelines to rectify the oversights of initial experiments  
59 [7][9][10].

60 Bayesian methods have been widely used in many areas of bioinformatics and proteomics mainly  
61 due to the fact that they lend themselves nicely to the challenge of analyzing complex, noisy and  
62 often incomplete data [11]. Their growing popularity over the last 20 years is mainly attributable to

63 advances in computational power which make fitting Bayesian models much more attainable for  
64 large datasets [12]. This article will review the literature on Bayesian methods in proteomics in  
65 general before focusing on how Bayesian methods can be used for the statistical analysis of mass  
66 spectra data for proteomic biomarker discovery. We will discuss the benefits of using Bayesian  
67 models compared to other traditional and machine learning methods and will also identify some  
68 reasons why Bayesian models might attain superior performance in the validation of separate  
69 cohorts. We also highlight methods used in other areas of research and other recent developments  
70 in Bayesian analysis which could prove to be useful in future applications of proteomic biomarker  
71 discovery experiments. Section 5.4 includes a worked proteomic discovery example for the  
72 prediction of cardiovascular disease where two of these methods are tested and compared against  
73 each other. This section also includes code samples which the reader can use to run these models  
74 for their own experiments using freely available software. Readers not interested in following the  
75 tutorial may wish to skip Section 5.4 and proceed directly to Section 6.

## 76 **2 What is Bayesian Inference?**

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

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

78 often expressed in words as:

79 *posterior is proportional to prior times likelihood*

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

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

94 Often the full form of the posterior distribution is unavailable due to the calculation of the  
95 normalising constant in the proportionality constraint. This problem is neatly sidestepped by using  
96 fitting methods such as Markov Chain Monte Carlo (MCMC) which make inference about the  
97 posterior distribution by sampling from it rather than computing it explicitly.

98

### 99 **3 Motivation for using Bayesian Methods**

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

108 One of the main reasons for the failure of many initial discovery studies to validate according to [14]  
109 is the failure to accurately model sources of experimental and biological variability. Many traditional  
110 pre-existing techniques have been used to analyse the data resulting from proteomic biomarker  
111 experiments such as support vector machines, random forests, lasso regression and various other  
112 classification methods [15][16][17]. However, the one disadvantage common to all these methods is  
113 that they ignore the uncertainty introduced to the data and assume that the experimental data are  
114 the only data available. This uncertainty can however be modelled and incorporated into a Bayesian  
115 framework in a consistent manner [18].

116 A common feature of proteomic discovery datasets is that the number of variables  $p$  tends to be  
117 much larger than the number of samples  $n$ , giving rise to problems of 'over-fitting' when traditional  
118 methods are used [19]. Over-fitting means that the chosen model fits the current data set too  
119 precisely, giving over-optimistic estimates of model performance that would not be repeated on an  
120 external validation cohort. The model is thus apportioning signal to random noise rather than  
121 identifying a true underlying model. Traditionally over-fitting is discouraged during the model  
122 building phase by adding a penalty for the complexity of the model. This is known as regularisation.  
123 Bayesian models overcome over-fitting in a similar manner, though the penalty is more explicitly  
124 stated via the prior distribution. Further flexibility can be obtained by marginalising over (i.e.  
125 removing through integration) or shrinking parameters [20][21] and so when used correctly will have  
126 a better chance of validating on a separate cohort. For example Kushner *et al* used a Bayesian Belief  
127 Network (BN) on both simulated and authentic proteomic data to discriminate between patients  
128 with sub types of Human T-cell Leukemia Virus type 1, and found that a BN with informative priors  
129 far outperformed traditional Linear and Quadratic Discriminant analysis with regards to cross  
130 validated and test accuracy [22]. They also found that biomarkers selected by the BN were far more  
131 stable over multiple iterations than the other methods tested. Similarly, Vannucci, Shaw and Brown  
132 [23] used probit models with Bayesian mixture priors and latent variables to classify women with  
133 ovarian cancer from their mass spectrum profiles and found that their method performed accurately  
134 and selected biomarker panels which were consistent with the literature.

### 135 **4 Bayesian models currently used in Proteomics**

#### 136 **4.1 Biomarker Discovery**

137 The objective of many biomarker discovery experiments is twofold; first to accurately classify  
138 samples into groups and second to select a subset of predictive peptide features or proteins which  
139 can further be validated and measured using specific targeted assays. Hence the analysis of

140 biomarker discovery data is not only a classification prediction problem but could also be viewed as  
141 a variable selection problem for high dimensional data where the number of important parameters  
142 is small.

143 With respect to proteomic mass spectrometric biomarker discovery, Bayesian models have not been  
144 as widely adopted as in other areas of proteomics. Some examples of Bayesian feature selection  
145 techniques have however emerged in the proteomics literature [22][23]. Yu and Chen [24] use an  
146 ovarian cancer dataset from the National Cancer Institute to showcase their proposed version of a  
147 hierarchical Bayesian neural network. They essentially deal with high dimensionality by filtering  
148 variables through the use of a Kolmogorov-Smirnov test which is used to set the hyper-parameters  
149 on the prior distribution of a variable being selected to a model. Deng, Geng and Ali [25] also  
150 proposed an interesting application of a Bayesian network where they used both microarray and  
151 mass spec experiments to choose biomarkers of prostate cancer. In this way their algorithm not only  
152 chose biomarkers which reported high predictive accuracy but also those which were supported by  
153 multiple sources of biological information. More recently the work of Serang *et al* used a Bayesian  
154 goodness of fit approach to detect the true number of differential features in an LC-MS experiment  
155 [26]. This method has the advantage that it avoids the need to specify many of the arbitrary cut-off  
156 choices common in most proteomic analyses such as the cut off for a “significant” fold change or a  
157 “significant” q value. With respect to isobaric labelled mass spectrometry data, Jow, Boys and  
158 Wilkinson proposed a hierarchical Bayesian method which was found to perform well in a variety of  
159 simulated and real proteomic experiments [27]. This Bayesian methodology in particular has the  
160 advantage that it can easily integrate multiple data experiments into a single model. Koh *et al*  
161 proposed an alternative model based approach for analysing data resulting from label based  
162 proteomic experiments [28]. They note that with labeling approaches many of the ratios used to  
163 identify differentially expressed proteins ignore the fact that some proteins are quantified using  
164 more peptides and are therefore more reproducible than others. Their proposed method  
165 incorporates this knowledge regarding the reproducibility of the protein quantities and models the  
166 hierarchical relationship between peptides and proteins directly giving greater importance to  
167 proteins which are the most reproducible.

## 168 **4.2 Other areas of Proteomics**

169 Bayesian methods have infiltrated many other areas of proteomic research apart from biomarker  
170 discovery. At a functional and structural level there are many examples of early adoptions to a  
171 probabilistic framework in areas such as sequence alignment [29][30] and predicting protein  
172 structures [31][32][33]. Bayesian methods have also proven popular for finding protein functions as  
173 well as predicting protein-protein interactions [34][35].

174 Mass spectrometry, and in particular peptide and protein identification from mass spectrum  
175 fragmentation data has also seen many contributions from Bayesian models. Serang, MacCross and  
176 Stafford Noble [36] proposed a Bayesian framework for protein inference with degenerate peptides  
177 to calculate the posterior probability of a given peptide belonging to a protein, which was found to  
178 outperform the popular software ProteinProphet on a number of datasets. Li *et al* [37] also  
179 proposed a fully probabilistic approach to protein identification, however their method was found to  
180 be too computationally intensive for use on large datasets [36]. ProteinProphet itself uses an  
181 empirical Bayes approach in which the prior is estimated from the data. Their method uses a two  
182 component mixture model to understand the distribution of peptide search scores observed for all

183 designated peptides [38][39]. Another software ProFound identifies proteins by searching through  
184 existing sequence databases and uses Bayes' theorem to calculate the posterior probability that  
185 each protein in the database is the current sample protein being analysed, given the experimental  
186 data and other available background information. Proteins are then ranked according to their  
187 posterior probability [40].

## 188 **5 Possible Bayesian applications for biomarker discovery**

189 The use of Bayesian methods for the statistical analysis of proteomic mass spec discovery data is still  
190 quite a new and emerging area and to-date has not reached the maturity of other areas of  
191 bioinformatics and systems biology. As mentioned previously, mass spectrometry discovery data  
192 tends to have far more variables ( $p$ ) than samples ( $n$ ) (commonly referred to as small  $n$  large  $p$ )  
193 however it is expected that very few of the variables measured are truly related to the response or  
194 outcome. For this reason many studies have focused on finding a small biomarker panel which can  
195 accurately predict disease [5][41]. There is a vast literature on Bayesian models used for feature  
196 selection on small  $n$  large  $p$  datasets in other areas of research, which thus could be applied to  
197 proteomic biomarker discovery. This section will outline some existing Bayesian models from the  
198 statistical literature which could have interesting applications to the area of proteomic biomarker  
199 discovery.

### 200 **5.1 Bayesian Lasso**

201 The Lasso model for linear regression is one of the most widely used methods for variable selection  
202 in high dimensional data [42]. The Lasso has also proven popular in various areas of proteomics.  
203 Huang et al [43] proposed ProteinLasso which uses the Lasso for protein inference; Friedman, Hastie  
204 and Tibshirani also showcased the graphical Lasso to find protein networks in flow cytometry cell  
205 signaling data [44]. For other examples see [45][46]. As previously stated, regularisation is a popular  
206 form of variable selection where a penalty is applied to the parameters in order to discourage  
207 complex models where many variables are chosen. The Lasso is a regularised version of ordinary  
208 least squares regression (for a continuous response) which balances model fit and model complexity  
209 by adding a penalty parameter which controls the absolute sum of the regression coefficients  
210 included in the model. The higher the penalty the more coefficients will have a value of zero and will  
211 be effectively eliminated from the model.

212 It was noticed that the coefficients returned by the original Lasso correspond to the mode of the  
213 posterior distribution in a Bayesian setting when a Laplace (double exponential) prior is placed on  
214 the parameter vector  $p(\beta|\sigma^2)$  where  $\sigma^2$  refers to the model variance; which led to the inception of  
215 the Bayesian Lasso [47]. The Bayesian Lasso allows for the full posterior of the model coefficients to  
216 be explored rather than just a point estimate, and so can give more instructive information  
217 regarding variable selection. Also, tuning parameters which control how harsh a penalty is placed on  
218 the model coefficients can be treated as unknown random variables and so their posterior  
219 distribution can also be sampled. This avoids the need for ad-hoc choices of tuning parameters such  
220 as those used in the traditional Lasso model. The Bayesian Lasso has been used in various  
221 biomedical and bioinformatics studies in recent years and has proven a popular approach for  
222 variable selection in data which have a sparse parameter space [48][49][50]. There is a freely  
223 available R package to run the Bayesian Lasso called "reglogit" [51] which is showcased in Section

224 5.4. The Bayesian Lasso can also be run using a package called rJAGS [52] which is shown in Section  
225 5.4.

## 226 5.2 Other Priors for Variable Selection

227 There is a wide array of literature proposing different shrinkage priors other than that of the Laplace  
228 prior (used by the Bayesian Lasso above) which have been shown to be optimal in various settings.  
229 For example Dunson DB and Lee suggest use of a generalized double Pareto prior [53] and Griffin  
230 and Brown suggest a normal-gamma prior on  $\beta$ , which is a generalisation of the double exponential  
231 prior [54]. One of the most popular rivals to the Laplace prior of the Bayesian Lasso is the Horseshoe  
232 prior [55]. Carvalho, Polson and Scott claim that the main advantages to the use of the Horseshoe  
233 prior is that it is robust to large signals and is very effective in shrinking noise variables [55]. The  
234 horseshoe distribution is very heavy tailed with an infinitely large spike at zero. This means that  
235 coefficients near zero can be shrunk very efficiently but also that coefficients far from zero will not  
236 be shrunk as severely, allowing for large signal if it is evident in the data. The horseshoe prior,  
237 Bayesian Lasso and ridge regression can be run using package “monomvn” in R [56].

238 A less severe shrinkage prior, though still widely used for Bayesian variable selection, is Zellner’s  $g$   
239 prior, where the prior on the parameter vector takes the form:  $\beta \sim N(\beta_0, g\sigma^2(X^T X)^{-1})$ . Here  $\beta_0$  is  
240 the value around which the regression coefficients are thought to centre (usually taken to be 0); the  
241 prior on  $\sigma^2$  is generally a non-informative prior and  $g$  is the hyper-parameter on the model  
242 coefficients controlling the degree of shrinkage [57]. This is a popular prior because of its  
243 computational simplicity for calculating marginal likelihoods (the likelihood function where some  
244 parameters have been removed through integration), and the fact that only  $g$  has to be estimated.  
245 There have been many suggestions on how to treat  $g$ . Some authors suggest placing a prior  
246 distribution whereas others suggest using fixed values or estimating the value for  $g$  using Empirical  
247 Bayes methods; see [58][59] for more information. Use of the  $g$  prior with a probit model has  
248 previously been proposed in the context of gene microarray studies to classify a number of diseases  
249 including colon cancer and leukemia [60] and also in gene selection for expression data [61], as well  
250 as in a ridge regression for high dimensional microarray breast cancer data [59]. Regression using the  
251  $g$  prior can be run in R using the BMS package [62].

252 An alternative to shrinkage is to directly model the probability of inclusion of a variable. A popular  
253 version of this for regression problems is that of Kuo and Mallick [63]. They introduce a vector of  
254 indicator variables which signify inclusion or exclusion of each parameter in the model as shown  
255 below [63].

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j I_j x_{ij} + \varepsilon_i$$

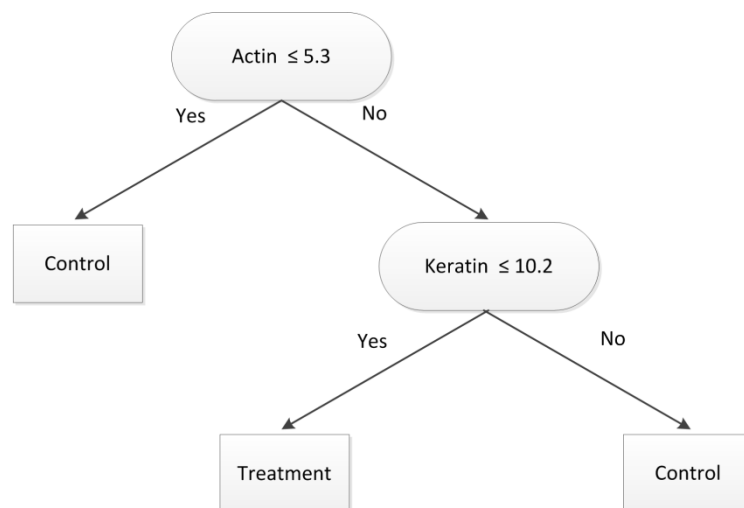
256 Here  $y_i$  is the value of response variable  $y$  for observation  $i$ ,  $\beta_j$  is the coefficient for variable  $j$ ,  $I_j$  is  
257 the indicator term which is 1 if variable  $j$  is to be included in the model and 0 otherwise,  $x_{ij}$  is the  
258 value of explanatory variable  $j$  for observation  $i$  and  $\varepsilon_i$  is random noise associated with observation  
259  $i$ . Usually independent priors are placed on the  $\beta$  vector and the indicator vector  $I$ . One of the  
260 advantages of this method is that it doesn’t require much tuning. However it can be slow to fit [64].

## 261 5.3 Bayesian non-parametric models

262 Both the Bayesian Lasso and other shrinkage priors used for regression models assume that the  
263 variables are linearly related and that variable interactions are known in the model specification. In  
264 this section we discuss examples of non-parametric Bayesian models which could be applied to  
265 proteomic biomarker discovery.

### 266 5.3.1 Bayesian CART

267 Classification and Regression decision trees (CART) are a popular method used in many areas of  
268 proteomics and in particular biomarker discovery [65][66][67] largely due to the fact they do not  
269 assume the covariates are linearly related to the response, often an unreasonably strong assumption  
270 in complex biological data. A key aspect of decision trees is that, unlike other regression models,  
271 they automatically select and include important variable interactions as part of the model building  
272 process and do not require these interactions to be explicitly specified in the model building.  
273 Decision trees consist of internal nodes where questions are asked based on a split rule consisting of  
274 a variable and a threshold, and terminal nodes which separate the observations into distinct groups.  
275 Observations which satisfy the split rule at each internal node are sent to the left hand daughter  
276 node and those which do not are sent to the right hand daughter node. Observations are iteratively  
277 split into left and right hand daughter nodes as they pass through each internal node in turn until a  
278 terminal node is reached. Figure 1 shows an example of a simple decision tree for a binary  
279 classification problem. Using Figure 1 as an example model to predict between two groups (control  
280 and treatment) it can be seen that all patients with Actin levels  $\leq 5.3$  would be predicted as  
281 belonging to the control group; all patients with Actin levels  $> 5.3$  and Keratin levels  $\leq 10.2$  would be  
282 predicted as treatment and all patients with Actin levels  $> 5.3$  and Keratin levels  $> 10.2$  would be  
283 predicted as control. As can be seen, decision trees are easily interpretable, naturally perform  
284 variable selection, allow for complex non-linear interactions and also perform prediction, which are  
285 the main reasons for their popularity in biomarker discovery studies [65][66][67].



286

287 **Figure 1 – Example of a binary classification decision tree**

288 The Bayesian version of CART (Bayesian CART) has not, to our knowledge, yet been used in  
289 proteomic biomarker discovery but could have very interesting applications in this area as they  
290 combine the advantages of Bayesian models with those of traditional decision trees discussed  
291 above. For a Bayesian CART model, the data in each terminal node of the tree is assumed to follow a  
292 multinomial distribution for classification problems. This allows the probability of belonging to a

293 given class to be computed which can provide richer information than merely knowing the predicted  
294 labels.

295 Chipman, George and McCulloch proposed a version of Bayesian CART which samples from the  
296 posterior distribution of the trees using MCMC [68]. This essentially means that many trees are  
297 sampled from the posterior distribution by creating a chain of  $k$  trees  $T_0, T_1 \dots T_k$ . At each iteration  
298  $k$ , a new tree is proposed and either accepted or rejected according to how well it matches the  
299 observed data and the prior distribution. The new trees are proposed by either growing or pruning  
300 (making the most recently accepted tree bigger or smaller) or changing or swapping internal nodes  
301 of the most recently accepted tree. This algorithm continues iteratively sampling trees until the  
302 model parameter estimates become stable. At this stage it is usually assumed that the algorithm has  
303 converged and we have samples from the posterior distribution of trees.

304

305 The tree size and shape is determined by a prior probability of a terminal node splitting  $P_{split} =$   
306  $\alpha (1 + d_i)^{-\beta}$  where  $d$  refers to the depth of the current terminal node  $i$  and  $\alpha$  and  $\beta$  are penalties  
307 on the tree size and shape respectively. As discussed, trees are sampled by growing, pruning,  
308 changing or swapping the current tree in the algorithm. Grow and prune moves are synonymous  
309 with node birth and death where a random terminal node is either converted into an internal node  
310 by further splitting it into two daughter nodes (grow) or a random internal node with two terminal  
311 daughter nodes is collapsed into a terminal node (prune). Growing or pruning a tree alters the size  
312 of the tree. Tree structure is altered by changing or swapping nodes where a new split rule is chosen  
313 for a given internal node in a tree or an internal parent-daughter pair of nodes is swapped around  
314 respectively.

### 315 5.3.2 Other Bayesian Tree Models

316 There have been many other variations on the Bayesian CART model of [68]. One application which  
317 was developed independently and at the same time is [69]. The Bayesian CART model of [69] differs  
318 from [68] in that their tree prior only requires one parameter determining the number of terminal  
319 nodes in the tree and does not include further parameters determining the tree shape and size. They  
320 also use a similar MCMC approach to sample from the posterior distribution of tree models [68].  
321 More recently [70] proposed an alternative prior on the tree  $\pi(T)$  as shown below:

$$\pi(T) = \alpha(m_0(T)) \prod_{u \in a(T)} \beta(m_{l(u)}(T) | m_u(T))$$

322 Here  $m_0(T)$  refers to the number of terminal nodes in tree  $T$ ,  $\alpha$  and  $\beta$  determine the tree size and  
323 shape respectively,  $m_u$  refers to the number of terminal nodes in the sub-tree below node  $u$  and  
324  $m_{l(u)}$  is the number of terminal left daughter nodes in the sub-tree below node  $u$ . In addition to the  
325 four proposal steps of [68] and [69] described above, [70] suggest an additional proposal move to  
326 sample new trees called the “restructure proposal”, this move searches for alternative trees that  
327 would result in the same terminal nodes as the current accepted tree. They claim that this  
328 overcomes the problem of slow mixing found in previous Bayesian CART models, as radically  
329 different trees can be proposed rather than just local changes of previously accepted trees.

### 330 5.3.3 Bayesian Additive Regression Trees (BART)

331 Bayesian Additive Regression Trees (BART) [71] is a fundamental extension of Bayesian CART and is a  
 332 non-parametric tree-based ensemble method which brings with it all the advantages of a fully  
 333 probabilistic model. The BART model uses a sum of multiple Bayesian CART trees as proposed in [68]  
 334 in its formulation. Posterior predictions are then constructed by adding the MCMC samples over all  
 335 trees. Point and credible intervals for each predicted data point can also be obtained by taking the  
 336 appropriate quantiles from the distribution of the MCMC samples for each predicted value of the  
 337 response variable.

338 The idea of CART ensemble methods is not new, in fact the most widely used of these methods, the  
 339 random forest, was invented in 2001 [72] and has been a popular method in the area of proteomic  
 340 biomarker discovery [41][73][74]. BART, to the authors' knowledge, has not yet been applied to a  
 341 proteomic setting, however a multiclass version has been used to classify satellite images [75] and a  
 342 very recent application to gene regulation data using informative priors for biomarker selection has  
 343 also been implemented [76]. Another extension of the BART method has also been proposed for use  
 344 on high dimensional survival data and was successfully showcased on a number of gene expression  
 345 datasets [77].

346 The recent addition of a more efficient parallelised software package in R called bartMachine [78]  
 347 means that this method can now be easily implemented for high dimensional datasets such as those  
 348 commonly found in MRM and smaller proteomic biomarker discovery experiments. Due to memory  
 349 constraints and computational complexity of the model however, some preprocessing of the data  
 350 may be needed for very high dimensional data. Code snippets to implement BART in R have been  
 351 included in the next section.

352 Bayesian CART and BART models offer many advantages over traditional tree and tree-ensemble  
 353 methods for biomarker discovery. Their main benefit is the fact that credible intervals can be  
 354 constructed around the point estimate of the probability of belonging to a given class. This could  
 355 have large implications for the quality of clinical decisions made based on the output of such  
 356 methods. Hence Bayesian tree models provide the user with a much richer output on which to make  
 357 decisions compared to traditional CART models.

358 A summary of the main advantages of the Bayesian Lasso, Bayesian Decision Tree methods and BART  
 359 discussed in this section can be seen in Table 1.

360 **Table 1:** Comparison of Bayesian Lasso model with Bayesian decision tree based models

	Bayesian Lasso	Bayesian Decision Tree	BART
Will work for small n large p data	✓	✓	✓
Automatically includes high order variable interactions		✓	✓
Eliminates non-predictive variables	✓		
Provides a variable importance score			✓

Finds linear relationships	✓	
Finds non-linear relationships	✓	✓

---

361

## 362 **5.4 Worked Example: Implementation of Bayesian Inference for biomarker** 363 **discovery in R**

364 This section will provide a worked example using the Bayesian Lasso and Bayesian additive  
 365 regression trees (BART) models described in the previous section using freely available software in  
 366 the statistical programming language R. Here we use an illustrative example of LC-MS data which  
 367 was collected for 500 patients, 150 of which had a cardiovascular disease and 350 of whom were  
 368 healthy. A total of 37 proteins were measured by MRM for each patient. The objective of this study  
 369 was two-fold as with any biomarker discovery experiment:

- 370 1. To build a classifier which can accurately predict between the two groups and
- 371 2. To find a subset of proteins/peptide features which are important in discriminating
- 372 between the groups

373 The dataset described is provided in the supplementary material where all identifying information  
 374 has been removed. Therefore peptide features in the following tutorial will be referred to using their  
 375 column number in the dataset provided rather than their sequence. This data set is shown for  
 376 illustrative purposes only, however the analyses and code used here are equivalent for higher  
 377 dimensional datasets where  $p \gg n$  as is common in shotgun discovery experiments. To compare fairly  
 378 across all analyses the data set was split into a training and test set where a random sample of 400  
 379 patients were chosen to build the model and the remaining 100 were used to test the model. This  
 380 ensures that none of the models over-fit to random artifacts in the data. In the following code  
 381 snippets the response variable for the training data will be referred to as “training\_response” and  
 382 the response variable for the test data as will be referred to as “test\_response”. Similarly, the  
 383 explanatory variables for the 400 training observations will be referred to “training\_data” and the  
 384 explanatory variables for the 100 test observations will be referred to as “test\_data”.

385 It should be noted that with shotgun proteomic experiments where the number of peptides  
 386 measured can reach tens of thousands much consideration should be given to the appropriate  
 387 sample size to use. Although sample size calculation is beyond the scope of this article the interested  
 388 reader may wish to refer to previous work where we show how varying sample sizes can affect the  
 389 overall classification performance of a model [79].

### 390 **5.4.1 Package reglogit**

391 An MCMC implementation of the Bayesian Lasso which is equivalent to a logistic regression with  
 392 double exponential priors can be run using the R package reglogit [51][80]. The Bayesian Lasso  
 393 model is run by default as follows:

```
394 1. set.seed(100)
395 2. #set number of iterations
396 3. T = 1000
397 4. reg_logit_model = reglogit(T, training_response, training_data, normalize=FALSE)
```

398 Simple prediction of the class of each sample can also be calculated via

```
399 1. reg_logit_preds = predict(reg_logit_model, XX=test_data)
```

400 The area under the ROC curve can be calculated as follows:

```
401 1. library(ROCR)
402 2. prediction=prediction(reg_logit_preds$mp,test_response)
403 3. performance=performance(prediction,"auc")
404 4. reg_logit_auc=performance@y.values[[1]][1]
```

405 and the classification rate can be calculated using the following code:

```
406 1. class_rate=sum(reg_logit_preds$c == test_response)
```

407 Analysis of the output for the cardiovascular disease data showed that the Bayesian Lasso model  
408 performed quite well in this instance and identified 74% of the test samples correctly with an area  
409 under the ROC curve of 0.65 (see Figure 2).

410 Important variables can be chosen by looking at the posterior distribution of the variable  
411 parameters. The 2.5%, 50% and 97.5% quantiles of the posterior distribution of the variable  
412 parameters can be viewed using the following code:

```
413 1. burnin = (1:(T/10))
414 2. quants = t(apply(reg_logit_model$beta[-
415 burnin,],2,function(x)quantile(x,probs=c(0.25,0.5,0.975))))
```

416 Choosing the variables whose 95% credible intervals did not include zero, we found that the  
417 peptides 3, 12 and 13 were given non-zero coefficients and hence were important in distinguishing  
418 between those patients who had experienced cardiovascular disease or not.

#### 419 5.4.2 JAGS

420 The Bayesian Lasso can also be run using a package called rjags [52] which requires installation of the  
421 JAGS software. The use of rjags does require some programming skills as all models are written by  
422 the user. However this allows the user full access to all information regarding the model and the  
423 ability to easily change priors and other model assumptions. An example of how a Bayesian Lasso  
424 might be run in JAGS is included in the following code:

```
425 1. require(rjags)
426 2. data=list(y=response,X=data,n=nrow(data),p=ncol(data))
427 3. #initialise parameter values
428 4. init=list(lambda=1,alpha=0,beta=rep(0,p))
429 5. modelstring="
430 model {
431   for (i in 1:n) {
432     #write a logistic regression
433     logit(theta[i])<-alpha+inprod(X[i,],beta)
434     y[i]~dbern(theta[i])
435   }
436   for (j in 1:p) {
437     #set a double exponential prior on the beta parameters (The Bayesian LASSO is
438     equivalent to placing a double exponential prior on beta)
439     beta[j]~ddexp(0,lambda)
440   }

```

```

441 16.     #set a vague prior on the intercept term. JAGS uses precision instead of varian
442       ce so this is equivalent to alpha~N(0,100)
443 17.     alpha~dnorm(0,0.01)
444 18.     #set a weakly informative prior on hyper-parameter lambda
445 19.     lambda~dgamma(0.1,0.1)
446 20.   }
447 21. "

```

448

449 Here non-informative priors have been set on the beta and lambda parameters, however these  
450 could be easily changed if previous experiments or expert opinion deemed some peptides to be  
451 better candidate biomarkers than others. The data used in this model includes all 500 samples  
452 where the response variable for the test samples has been set to NA. rjags will automatically give  
453 predicted values for all samples whose response is flagged as NA.

454 Once the model has been run, the model parameter output can be viewed using the following code:

```

455 1. #use the coda.samples command to view parameter output and diagnostics
456 2. Output=coda.samples(model=model,variable.names=c("alpha","beta","lambda"),n.iter=10
457    00,thin=10)

```

458 Convergence of the model parameters and the quantiles of the model parameters can also be  
459 checked:

```

460 1. #check convergence
461 2. gelman.diag(output)
462 3. gelman.plot(output)

```

463 The 95% credible interval for the parameter estimates can be viewed as follows:

```

464 1. #get quantiles of the model parameters
465 2. quantiles=summary(window(output,burnin=2000))[[2]]

```

466 Choosing those variable parameters whose 2.5% and 97.5% quantiles do not include 0 found that  
467 variables 3, 12 and 27 had non-zero coefficients according to this model, which corresponds to  
468 peptides numbered 3, 12 and 27 as being important predictors of cardiovascular disease. In this case  
469 the JAGS model predicted 67% of the test cases correctly and gave an area under the ROC curve of  
470 0.66 (see Figure 2).

471 As mentioned earlier, the posterior distribution of the model parameters can be explored in a  
472 Bayesian setting. This means that credible intervals for the probability of having a cardiovascular  
473 event can be constructed with these models rather than just having access to a point prediction.

474 The quantiles of the predicted probabilities for each of the test samples can be viewed as follows:

```

475 1. #see quantiles for theta
476 2. output_theta=coda.samples(model=model,variable.names=c("theta"),
477   n.iter=1000,thin=10)
478 4. #plot 95% CI for the probability of having cardiovascular disease
479 5. theta_quants=summary(output_theta[,test_samples]][[2]]

```

480 To illustrate the usefulness of this additional information we shall take two patients numbered 60  
481 and 367. Patient 60 has a median predicted probability of 25.61% of having cardiovascular disease

482 and patient 367 has a median predicted probability of 4.25%. In the absence of further information  
483 (as with machine learning methods) the clinician would give both patients the all clear. However if  
484 we look at the 2.5% and 97.5% quantiles for patient 60 we see they range between 7.46% and  
485 62.79% respectively whereas those for patient 367 range from 1.01% to 13.29%. This additional  
486 information means that the model is very sure that patient 367 does not have cardiovascular  
487 disease; however it is not at all sure as to the class of patient 60. If a clinician was merely basing the  
488 prognosis on the median estimate, they would most likely quite confidently diagnose both patients  
489 as healthy. However knowing that the estimate for patient 60 could vary anywhere between 7.46%  
490 and 62.79% might change their opinion and hence the medical advice offered to this patient. For  
491 example patient 60 may be sent for further diagnostic tests. Alternatively, knowing the 95% credible  
492 interval for patient 367 ranges from 1.01% to 13.29%, might give the clinician added confidence as  
493 to the true diagnosis of this patient. In reality patient 60 did have cardiovascular disease and patient  
494 367 was healthy, so basing predictions on a point estimate would have led to the misdiagnosis of  
495 patient 60 in this case. Bayesian decision theory [81], not discussed here, provides a more nuanced  
496 approach to making such decisions.

### 497 5.4.3 bartMachine

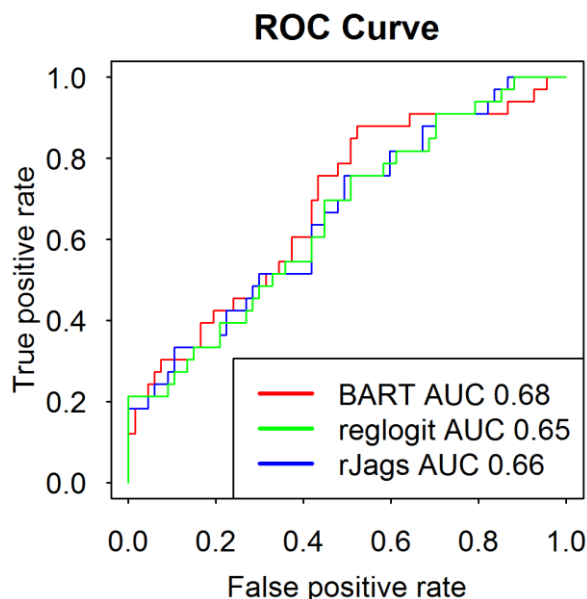
498 BART can be run using the package bartMachine in R [82]. The bartMachine function creates a Java  
499 virtual machine which requires the user to set the amount of memory to be used by the function.  
500 This can be set using the options() command in R as shown in the following code. Here we have set  
501 5Gb of RAM as the limit for the virtual machine. The serialize=TRUE command allows the BART  
502 model to be saved and loaded at a later date otherwise all information regarding the model will be  
503 lost once R is closed. The bartMachine model can be run on the training data as follows:

```
504 1. options(java.parameters = "-Xmx5000m")  
505 2. library(bartMachine)  
506 3. set.seed(100)  
507 4. bm=bartMachine(data.frame(training_data),training_response,num_trees=50,num_iterati  
508 ons_after_burn_in = 1000,mem_cache_for_speed = FALSE,serialize = TRUE)
```

509 Once the model is run on the training data using the code above, predictions and the full posterior of  
510 the test samples can be viewed as follows:

```
511 1. #get predicted values for test samples  
512 2. preds=bart_predict_for_test_data(bm,data.frame(scale(cvd_data[test_samps,])),respon  
513 se[test_samps])  
514 3. #look at the posterior samples for test data  
515 4. test_samps_posterior=bart_machine_get_posterior(bm,data.frame(scale(cvd_data[test_s  
516 amps,])))  
517 5. #get the 2.5%,50% and 97,5% quantiles of the posterior test samples  
518 6. posterior_quantiles=t(apply(test_samps_posterior[[3]],1,function(x)quantile(x,probs  
519 =c(0.025,0.5,0.975))))
```

520 In this case BART marginally outperformed both implementations of the Lasso for this data in terms  
521 of the classification rate of 71% for the test samples. Figure 2 shows the ROC curve for all three meth  
522 ods compared in this section. As can be seen BART also outperformed both implementations of the L  
523 asso in terms of the area under the ROC curve giving an AUC of 0.68.  
524



525  
526 **Figure 2** - ROC curve comparison of Lasso using package reglogit and rJags and BART using package  
527 bartMachine

528 bartMachine unlike the Lasso does not eliminate variables from the model and as such all variables a  
529 re given an importance score which is based on the number of times each variable was selected for e  
530 ach tree in the model. The importance score for each variable can be seen by using the following co  
531 mmand:

```
532 1. #get variable importance scores  
533 2. get_var_props_over_chain(bm)
```

534 In this case the bartMachine model found that variables 13, 2, 14, 3, 34, 23 and 12 were given the hi  
535 ghest importance. It is interesting to note that all but one of the variables identified by both Lasso im  
536 plementations were also identified as the most important according to BART. A vignette for the bart  
537 Machine package can be found at [78].

## 538 6 Discussion

539 In this paper we have identified a number of potential areas from the Bayesian statistical literature  
540 which could be applied to proteomic discovery data. We have also guided the reader towards using  
541 these models for their own experiments by supplying tutorial style example code using freely  
542 available software and have showcased these methods using a real proteomic experiment to predict  
543 cardiovascular disease.

544 Bayesian modeling for proteomic biomarker discovery data is a relatively new and emerging field  
545 with exciting future opportunities. Many articles have discussed reasons for the failure of biomarker  
546 panels from numerous biomarker discovery experiments to validate on separate cohorts. Some of  
547 the issues cited in the literature could be alleviated by the incorporation of a probabilistic model.  
548 One such reason suggested is that the original model over-fits the data. This is due to the fact that  
549 biomarker discovery datasets (especially those emerging from mass spectrometry) tend to be small  
550  $n$  large  $p$ . Bayesian models when used with sensible priors tend to be more robust to over-fitting  
551 than non-Bayesian models as the full posterior distribution of the model parameters is given.  
552 Another possible reason for the failure of many experiments to validate is the failure to incorporate

553 industry and scientific knowledge about underlying processes to the model. With careful modeling  
554 Bayesian methods can easily include information about variability introduced to the data through  
555 the various experimental and preprocessing stages before data are subjected to statistical analysis  
556 rather than ignore this variability.

557 The area of Bayesian variable selection for high dimensional data is a new and rapidly developing  
558 field. Decision tree models and ensemble methods in particular have traditionally been very popular  
559 for biomarker discovery as they are easily interpreted, are non parametric and automatically include  
560 important interactions without prior specification by the user. There have been many recent  
561 advances in Bayesian tree and ensemble methods which could have very interesting applications for  
562 proteomic biomarker discovery.

### 563 **Acknowledgements**

564 We would like to thank Dr Chris Watson and Dr John Baugh for kindly providing the cardiovascular  
565 dataset used in this manuscript. We would also like to acknowledge that this work was supported by  
566 the Irish Research Council. Protein biomarker discovery work in the Pennington Biomedical  
567 Proteomics Group is supported by grants from Science Foundation Ireland (for mass spectrometry  
568 instrumentation), the Irish Cancer Society (PCI11WAT), St Luke's Institute for Cancer Research, the  
569 Health Research Board (HRA\_POR/2011/125), Movember GAP and the EU FP7 (MIAMI). The UCD  
570 Conway Institute is supported by the Program for Research in Third Level Institutions as  
571 administered by the Higher Education Authority of Ireland.

572 **References**

- 573 [1] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics.  
574 *Bioinformatics* 2007;23:2507–17. doi:10.1093/bioinformatics/btm344.
- 575 [2] Chambers AG, Percy AJ, Yang J, Camenzind AG, Borchers CH. Multiplexed quantitation of  
576 endogenous proteins in dried blood spots by multiple reaction monitoring-mass  
577 spectrometry. *Mol Cell Proteomics* 2013;12:781–91. doi:10.1074/mcp.M112.022442.
- 578 [3] Mikolajczyk SD, Song Y, Wong JR, Matson RS, Rittenhouse HG. Are multiple markers the  
579 future of prostate cancer diagnostics? *Clin Biochem* 2004;37:519–28.  
580 doi:10.1016/j.clinbiochem.2004.05.016.
- 581 [4] Ky B, French B, Levy WC, Sweitzer NK, Fang JC, Wu AHB, et al. Multiple biomarkers for risk  
582 prediction in chronic heart failure. *Circ Heart Fail* 2012;5:183–90.  
583 doi:10.1161/CIRCHEARTFAILURE.111.965020.
- 584 [5] Surinova S, Schiess R, Hüttenhain R, Cerciello F, Wollscheid B, Aebersold R. On the  
585 development of plasma protein biomarkers. *J Proteome Res* 2011;10:5–16.  
586 doi:10.1021/pr1008515.
- 587 [6] Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation : the long and  
588 uncertain path to clinical utility 2006;24:971–83. doi:10.1038/nbt1235.
- 589 [7] Diamandis EP. The failure of protein cancer biomarkers to reach the clinic: why, and what can  
590 be done to address the problem? *BMC Med* 2012;10:87. doi:10.1186/1741-7015-10-87.
- 591 [8] Oon S, Pennington S, Fitzpatrick J, Watson RWG. Biomarker research in prostate cancer-  
592 towards utility, not futility. *Nat Rev Urol* 2011;8:131–8.
- 593 [9] Dakna M, Harris K, Kalousis A, Carpentier S, Kolch W, Schanstra JP, et al. Addressing the  
594 challenge of defining valid proteomic biomarkers and classifiers. *BMC Bioinformatics*  
595 2010;11:594. doi:10.1186/1471-2105-11-594.
- 596 [10] Alaiya A, Al-Mohanna M, Linder S. Clinical cancer proteomics: promises and pitfalls. *J*  
597 *Proteome Res* 2005;4:1213–22. doi:10.1021/pr050149f.
- 598 [11] Wilkinson DJ. Bayesian methods in bioinformatics and computational systems biology. *Brief*  
599 *Bioinform* 2007;8:109–16. doi:10.1093/bib/bbm007.
- 600 [12] Beaumont M a, Rannala B. The Bayesian revolution in genetics. *Nat Rev Genet* 2004;5:251–  
601 61. doi:10.1038/nrg1318.
- 602 [13] Box GEP, Tiao GC. *Bayesian Inference in Statistical Analysis*. Wiley Clas. Addison-Wesley Pub.  
603 Co.; 1973.
- 604 [14] Harris K, Girolami M, Mischak H. Definition of Valid Proteomic Biomarkers : A Bayesian  
605 Solution. In: Kadirkamanathan V, Sanguinetti G, Girolami M, Niranjana M, Noirel J, editors.  
606 *Pattern Recognit. Bioinformatics, Lect. Notes Comput. Sci.*, Springer Berlin Heidelberg; 2009,  
607 p. 137–49.

- 608 [15] Sampson DL, Parker TJ, Upton Z, Hurst CP. A Comparison of Methods for Classifying Clinical  
609 Samples Based on Proteomics Data: A Case Study for Statistical and Machine Learning  
610 Approaches. *PLoS One* 2011;6:e24973. doi:10.1371/journal.pone.0024973.
- 611 [16] Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, et al. Comparison of statistical  
612 methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*  
613 2003;19:1636–43. doi:10.1093/bioinformatics/btg210.
- 614 [17] Caruana R, Karampatziakis N, Yessenalina A. An empirical evaluation of supervised learning in  
615 high dimensions. *Proc. 25th Int. Conf. Mach. Learn. (ICML '08)*, 2008, p. 96–103.
- 616 [18] Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data:  
617 regularized t-test and statistical inferences of gene changes. *Bioinformatics* 2001;17:509–19.
- 618 [19] Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning Data  
619 Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics; 2009.
- 620 [20] Waterhouse S, MacKay D, Robinson T. *Advances in Neural Information Processing Systems 8*.  
621 In: Touretzky DS, Mozer MC, Hasselmo ME, editors. *Bayesian Methods Mix. Expert*. 8th ed.,  
622 MIT Press; 1995, p. 351–7.
- 623 [21] Bishop CM, Tipping ME. *Bayesian Regression and Classification*. *NATO Sci Ser III Computer Syst  
624 Sci n.d.*;190:267–88.
- 625 [22] Kuschner KW, Malyarenko DI, Cooke WE, Cazares LH, Semmes OJ, Tracy ER. A Bayesian  
626 network approach to feature selection in mass spectrometry data. *BMC Bioinformatics*  
627 2010;11:177. doi:10.1186/1471-2105-11-177.
- 628 [23] Vannucci M, Sha N, Brown PJ. NIR and mass spectra classification: Bayesian methods for  
629 wavelet-based feature selection. *Chemom Intell Lab Syst* 2005;77:139–48.  
630 doi:10.1016/j.chemolab.2004.10.009.
- 631 [24] Yu J, Chen X-W. Bayesian neural network approaches to ovarian cancer identification from  
632 high-resolution mass spectrometry data. *Bioinformatics* 2005;21 Suppl 1:i487–94.  
633 doi:10.1093/bioinformatics/bti1030.
- 634 [25] Deng X, Geng H, Ali HH. Cross-platform analysis of cancer biomarkers: a Bayesian network  
635 approach to incorporating mass spectrometry and microarray data. *Cancer Inform*  
636 2007;3:183–202.
- 637 [26] Serang O, States U, Steen J a. Nonparametric Bayesian Evaluation of Differential Protein  
638 Quantification. *J Proteome Res* 2013;12:4556–65.
- 639 [27] Jow H, Boys RJ, Wilkinson DJ. Bayesian identification of protein differential expression in  
640 multi-group isobaric labelled mass spectrometry data. *Stat Appl Genet Mol Biol* 2014;13:531–  
641 51. doi:10.1515/sagmb-2012-0066.
- 642 [28] Koh HW., Swa HLF, Damian F, Ler SG, Gunaratne J, Choi H. EBprot: Statistical Analysis of  
643 Labeling-based Quantitative Proteomics Data. *Proteomics* 2015.  
644 doi:10.1002/pmic.201400620.

- 645 [29] Liu JS, Neuwald, Andrew F. Lawrence CE. Bayesian Models for Multiple Local Sequence  
646 Alignment and Gibbs Sampling Strategies. *J Am Stat Assoc* 1995;90:1156–70.  
647 doi:10.1080/01621459.1995.10476622.
- 648 [30] Webb B-JM, Liu JS, Lawrence CE. BALSAs: Bayesian algorithm for local sequence alignment.  
649 *Nucleic Acids Res* 2002;30:1268–77.
- 650 [31] Schmidler SC, Liu JS, Brutlag DL. Bayesian Segmentation of Protein Secondary Structure. *J*  
651 *Comput Biol* 2000;7:233–48. doi:10.1089/10665270050081496.
- 652 [32] Brevern AG De, Etchebest C, Hazout S. Bayesian Probabilistic Approach for Predicting  
653 Backbone Structures in Terms of Protein Blocks. *ProteinsStructure,Function Genet*  
654 *2000;3:271–87.*
- 655 [33] Aydin Z, Singh A, Bilmes J, Noble WS. Learning sparse models for a dynamic Bayesian network  
656 classifier of protein secondary structure. *BMC Bioinformatics* 2011;12:154. doi:10.1186/1471-  
657 2105-12-154.
- 658 [34] Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of Protein Function Using Protein–  
659 Protein Interaction Data. *J Comput Biol* 2003;10:947–60. doi:10.1089/10665270322756168.
- 660 [35] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, et al. A Bayesian networks  
661 approach for predicting protein-protein interactions from genomic data. *Science*  
662 *2003;302:449–53.* doi:10.1126/science.1087361.
- 663 [36] Serang O, MacCross MJ, Stafford Noble W. Efficient marginalization to compute protein  
664 posterior probabilities from shotgun mass spectrometry data. *J Proteome Res* 2010;9:5346–  
665 57. doi:10.1021/pr100594k.Efficient.
- 666 [37] Li YF, Arnold RJ, Li Y, Radivojac P, Sheng Q, Tang H. A bayesian approach to protein inference  
667 problem in shotgun proteomics. *J Comput Biol* 2009;16:1183–93.  
668 doi:10.1089/cmb.2009.0018.
- 669 [38] Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A Statistical Model for Identifying Proteins by  
670 Tandem Mass Spectrometry abilities that proteins are present in a sample on the basis. *Anal*  
671 *Chem* 2003;75:4646–58.
- 672 [39] Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by  
673 tandem mass spectrometry. *Nat Methods* 2007;4:787–97. doi:10.1038/NMETH1088.
- 674 [40] Zhang W, Chait BT. ProFound : An Expert System for Protein Identification Using Mass  
675 Spectrometric Peptide Mapping Information. *Anal Chem* 2000;72:2482–9.
- 676 [41] Cima I, Schiess R, Wild P, Kaelin M, Schuffler P, Lange V, et al. Cancer genetics-guided  
677 discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. *Proc*  
678 *Natl Acad Sci* 2011:1–6. doi:10.1073/pnas.1013699108.
- 679 [42] Tibshirani R. lasso.pdf. *J R Stat Soc Ser B (Statistical Methodol* 1996;58:267–88.

- 680 [43] Huang T, Gong H, Yang C, He Z. ProteinLasso: A Lasso regression approach to protein  
681 inference problem in shotgun proteomics. *Comput Biol Chem* 2013;43:46–54.  
682 doi:10.1016/j.compbiolchem.2012.12.008.
- 683 [44] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical  
684 lasso. *Biostatistics* 2008;9:432–41. doi:10.1093/biostatistics/kxm045.
- 685 [45] Qin T, Tsoi LC, Sims KJ, Lu X, Zheng WJ. Signaling network prediction by the Ontology  
686 Fingerprint enhanced Bayesian network. *BMC Syst Biol* 2012;6 (Suppl S. doi:10.1186/1752-  
687 0509-6-S3-S3.
- 688 [46] Soltys SG, Le Q, Shi G, Tibshirani R, Giaccia AJ, Koong AC. The Use of Plasma Surface-  
689 Enhanced Laser Desorption / Ionization Time-of-Flight Mass Spectrometry Proteomic Patterns  
690 for Detection of Head and Neck Squamous Cell Cancers The Use of Plasma Surface-Enhanced  
691 Laser Desorption / Ionization Time-of-Flight Mass. *Clin Cancer Res* 2004;10:4806–12.
- 692 [47] Park T, Casella G. The Bayesian Lasso. *J Am Stat Assoc* 2008;103:681–6.  
693 doi:10.1198/016214508000000337.
- 694 [48] Li J, Das K, Fu G, Li R, Wu R. The Bayesian lasso for genome-wide association studies.  
695 *Bioinformatics* 2011;27:516–23. doi:10.1093/bioinformatics/btq688.
- 696 [49] De los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, et al. Predicting  
697 quantitative traits with regression models for dense molecular markers and pedigree.  
698 *Genetics* 2009;182:375–85. doi:10.1534/genetics.109.101501.
- 699 [50] Cai X, Huang A, Xu S. Fast empirical Bayesian LASSO for multiple quantitative trait locus  
700 mapping. *BMC Bioinformatics* 2011;12:211. doi:10.1186/1471-2105-12-211.
- 701 [51] Gramacy RB, Polson NG. Simulation-based regularized logistic regression 2014.  
702 doi:10.1214/12-BA719.
- 703 [52] Stukalov A, Plummer MM. Package “rjags” 2015. [http://cran.r-](http://cran.r-project.org/web/packages/rjags/rjags.pdf)  
704 [project.org/web/packages/rjags/rjags.pdf](http://cran.r-project.org/web/packages/rjags/rjags.pdf) (accessed February 12, 2015).
- 705 [53] Armagan A, Dunson DB, Lee J. Generalized Double Pareto Shrinkage. *Stat Sin* 2013;23:119–  
706 43.
- 707 [54] Griffin JE, Brown PJ. Inference with normal-gamma prior distributions in regression problems.  
708 *Bayesian Anal* 2010;5:171–88. doi:10.1214/10-BA507.
- 709 [55] Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. *Biometrika*  
710 2010;97:465–80. doi:10.1093/biomet/asq017.
- 711 [56] Gramacy RB. Package monomvn 2014. [http://cran.r-](http://cran.r-project.org/web/packages/monomvn/monomvn.pdf)  
712 [project.org/web/packages/monomvn/monomvn.pdf](http://cran.r-project.org/web/packages/monomvn/monomvn.pdf) (accessed September 12, 2014).
- 713 [57] Zellner A. On assessing prior distributions and Bayesian regression analysis with g-prior  
714 distributions. In: P.K G, Zellner A, editors. *Bayesian Inference Decis. Tech. Essays Honor Bruno*  
715 *Finetti*, North-Holland/Elsevier; 1986, p. 233–43.

- 716 [58] Liang F, Paulo R, Molina G, Clyde MA, Berger JO. Mixtures of  $g$ -priors for Bayesian Variable  
717 Selection. *J Am Stat Assoc* 2008;103:410–23. doi:10.1198/016214507000001337.
- 718 [59] Baragatti M, Pommeret D. A study of variable selection using  $g$ -prior distribution with ridge  
719 parameter. *Comput Stat Data Anal* 2012;56:1920–34. doi:10.1016/j.csda.2011.11.017.
- 720 [60] Yang A-J, Song X-Y. Bayesian variable selection for disease classification using gene expression  
721 data. *Bioinformatics* 2010;26:215–22. doi:10.1093/bioinformatics/btp638.
- 722 [61] Zhou X, Wang X, R. Dougherty E. A Bayesian approach to nonlinear probit gene selection and  
723 classification. *J Franklin Inst* 2004;341:137–56. doi:10.1016/j.jfranklin.2003.12.010.
- 724 [62] Feldkircher, Martin Zeugner S. Package BMS 2013. [http://cran.r-](http://cran.r-project.org/web/packages/BMS/BMS.pdf)  
725 [project.org/web/packages/BMS/BMS.pdf](http://cran.r-project.org/web/packages/BMS/BMS.pdf) (accessed September 12, 2014).
- 726 [63] Kuo L, Mallick B. Variable Selection for Regression Models. *Indian J Stat Spec Issue Bayesian*  
727 *Anal* 1998;60:65–81.
- 728 [64] O’Hara RB, Sillanpää MJ. A review of Bayesian variable selection methods: what, how and  
729 which. *Bayesian Anal* 2009;4:85–118. doi:10.1214/09-BA403.
- 730 [65] Vlahou A, Schorge JO, Gregory BW, Coleman RL. Diagnosis of Ovarian Cancer Using Decision  
731 Tree Classification of Mass Spectral Data. *J Biomed Biotechnol* 2003;2003:308–14.  
732 doi:10.1155/S1110724303210032.
- 733 [66] Yu Y, Chen S, Wang L-S, Chen W-L, Guo W-J, Yan H, et al. Prediction of pancreatic cancer by  
734 serum biomarkers using surface-enhanced laser desorption/ionization-based decision tree  
735 classification. *Oncology* 2005;68:79–86. doi:10.1159/000084824.
- 736 [67] Markey MK, Tourassi GD, Floyd CE. Decision tree classification of proteins identified by mass  
737 spectrometry of blood serum samples from people with and without lung cancer. *Proteomics*  
738 2003;3:1678–9. doi:10.1002/pmic.200300521.
- 739 [68] Chipman H a., George EI, McCulloch RE. Bayesian CART Model Search. *J Am Stat Assoc*  
740 1998;93:935–60.
- 741 [69] Denison BYDGT, Mallick BANK, Smith AFM. A Bayesian CART algorithm. *Biometrika*  
742 1998;85:363–77.
- 743 [70] Wu Y, Tjelmeland H, West M. Bayesian CART – Prior Specification and Posterior Simulation –.  
744 *J Comput Graph Stat* 2006;16:44–66.
- 745 [71] Chipman H a., George EI, McCulloch RE. BART: Bayesian additive regression trees. *Ann Appl*  
746 *Stat* 2010;4:266–98. doi:10.1214/09-AOAS285.
- 747 [72] Breiman L. Random forests. *Mach Learn* 2001;45:5–32. doi:10.1186/1478-7954-9-29.
- 748 [73] Fan Y, Murphy TB, Byrne JC, Brennan L, Fitzpatrick JM, Watson RWG. Applying random forests  
749 to identify biomarker panels in serum 2D-DIGE data for the detection and staging of prostate  
750 cancer. *J Proteome Res* 2011;10:1361–73. doi:10.1021/pr1011069.

- 751 [74] Baek S, Tsai C-A, Chen JJ. Development of biomarker classifiers from high-dimensional data.  
752 Brief Bioinform 2009;10:537–46. doi:10.1093/bib/bbp016.
- 753 [75] Agarwal R, Ranjan P, Chipman H. A new Bayesian ensemble of trees approach for land cover  
754 classification of satellite imagery. Can J Remote Sens 2013;39:507–20. doi:10.5589/m14-003.
- 755 [76] Bleich J, Kapelner A, George EI, Jensen ST. Variable Selection for BART : An Application to  
756 Gene Regulation. Ann Appl Stat 2014;8:1750–81.
- 757 [77] Bonato V, Baladandayuthapani V, Broom BM, Sulman EP, Aldape KD, Do K-A. Bayesian  
758 ensemble methods for survival prediction in gene expression data. Bioinformatics  
759 2011;27:359–67. doi:10.1093/bioinformatics/btq660.
- 760 [78] Kapelner A, Bleich J. bartMachine: Machine Learning With Bayesian Additive Regression Trees  
761 2013. <http://cran.r-project.org/web/packages/bartMachine/vignettes/bartMachine.pdf>  
762 (accessed March 2, 2015).
- 763 [79] Hernández B, Parnell A, Pennington SR. Why have so few proteomic biomarkers “survived”  
764 validation? (Sample size and independent validation considerations). Proteomics  
765 2014;14:1587–92. doi:10.1002/pmic.201300377.
- 766 [80] Gramacy RB, Polson NG. Simulation-based regularized logistic regression. Bayesian Anal  
767 2012;7:567–90. doi:10.1214/12-BA719.
- 768 [81] Berger JO. Statistical decision theory and Bayesian analysis. 2nd ed. New York: Springer;  
769 1985.
- 770 [82] Kapelner A, Bleich J. Package bartMachine 2014. [http://cran.r-](http://cran.r-project.org/web/packages/bartMachine/bartMachine.pdf)  
771 [project.org/web/packages/bartMachine/bartMachine.pdf](http://cran.r-project.org/web/packages/bartMachine/bartMachine.pdf).
- 772
- 773

774 **Figure Legends:**

775 **Figure 1 – Example of a binary classification decision tree**

776 **Figure 2 – ROC curve comparison of Lasso using package reglogit and rJags and BART using package**  
777 **bartMachine**

778 **Table Captions:**

779 **Table 1 – Comparison of Bayesian Lasso model with Bayesian decision tree based models**

780

781

782