VIEWPOINT

Why have so few proteomic biomarkers "survived" validation? (Sample size and independent validation considerations)

Belinda Hernández[1,2], *Andrew Parnell*[1] *and Stephen R. Pennington*[2]

[1] Complex and Adaptive Systems Laboratory, School of Mathematical Sciences (Statistics), University College Dublin, Dublin, Ireland
[2] School of Medicine and Medical Science, UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland

Proteomic biomarker discovery has led to the identification of numerous potential candidates for disease diagnosis, prognosis, and prediction of response to therapy. However, very few of these identified candidate biomarkers reach clinical validation and go on to be routinely used in clinical practice. One particular issue with biomarker discovery is the identification of significantly changing proteins in the initial discovery experiment that do not validate when subsequently tested on separate patient sample cohorts. Here, we seek to highlight some of the statistical challenges surrounding the analysis of LC-MS proteomic data for biomarker candidate discovery. We show that common statistical algorithms run on data with low sample sizes can overfit and yield misleading misclassification rates and AUC values. A common solution to this problem is to prefilter variables (via, e.g. ANOVA and or use of correction methods such as Bonferonni or false discovery rate) to give a smaller dataset and reduce the size of the apparent statistical challenge. However, we show that this exacerbates the problem yielding even higher performance metrics while reducing the predictive accuracy of the biomarker panel. To illustrate some of these limitations, we have run simulation analyses with known biomarkers. For our chosen algorithm (random forests), we show that the above problems are substantially reduced if a sufficient number of samples are analyzed and the data are not prefiltered. Our view is that LC-MS proteomic biomarker discovery data should be analyzed without prefiltering and that increasing the sample size in biomarker discovery experiments should be a very high priority.

Received: August 27, 2013
Revised: March 19, 2014
Accepted: April 7, 2014

Keywords:
Bioinformatics / Biomarker panels / Cross-validation / Proteomic discovery / Random forest / Sample size

1 Introduction

Biomarker discovery in proteomics has resulted in a very large number of publications describing potential biomarkers for the detection and prognosis of a large range of diseases. For example, a PubMed search for the terms "protein and biomarker discovery" undertaken on 19 February 2014 revealed a total of 6,690 publications in this area. However, it has been widely noted that this fervent discovery and publication of potential biomarkers has not translated to a comparable increase in the number of clinically accepted "proteomic" tests. In fact very few, if any, biomarkers discovered by "proteomics" are routinely used in a clinical setting despite large government and industry investment [1–3].

One of the main reasons for potential biomarkers failing to be used in clinical practice is that many are deemed significant in an initial discovery cohort but are later found not to be significant in subsequent validation studies [1]. In our opinion, this is usually due to a combination of model overfitting due to small samples sizes in the initial

Correspondence: Belinda Hernández, School of Medicine and Medical Science, UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield Campus, Dublin 4, Ireland
E-mail: belinda.hernandez@ucdconnect.ie

Abbreviations: AUC, area under the receiver operating curve; **RF**, random forest; **FDR**, false discovery rate discovery experiment and/or incorrect use of cross-validation. The serious consequences of model overfitting and incorrectly applied statistical techniques in "omic" biomarker discovery were highlighted last year in an Institute of Medicine review of research at Duke University, which resulted in a high number of manuscript retractions and the cancellation of three clinical trials [4]. Here, we illustrate our view by showing practical examples of how small sample sizes, model overfitting, and prefiltering can lead to deceptive results. Although a number of these mistakes have been noted previously (see [5–8]), it seems that as yet they are not well-known in the area of proteomic biomarker discovery. In this Viewpoint, we seek to raise awareness of some statistical challenges associated with proteomic biomarker discovery, and in so doing impact positively on future protein biomarker studies.

In recent years, LC-MS has become one of the leading approaches for peptide identification and quantification as it can reliably quantify and identify a large number of peptides in a reasonably short time. LC-MS data have been used in many biomarker discovery experiments (for examples see [9, 10]). Typical LC-MS experiments measure many tens of thousands of variables or m/z features (which we will now refer to as p) on a relatively small number of samples (hereon n; often much less than 100). The resulting data present what is often referred to as a "small n large p" problem. Technical aspects in generating and processing LC-MS have been well summarized in [11] and are not discussed here.

Likewise, the "small n large p" problem has been well documented. Many machine learning and statistical techniques have been identified that may be used to analyze such data, including RFs, regularized logistic regression, support vector machines, k-nearest neighbors, neural networks, and shrunken centroids. These and other methods have been reviewed and compared in a number of previous studies [5, 12, 13]. For brevity, we focus here solely on the random forest (RF) method as this is one of the fastest and most flexible techniques that is applied to biomarker discovery [13–15]. However, the issues that arise from data prefiltering and small sample sizes are common to all techniques.

RF was first proposed by Leo Breiman in 2001 [16] and has proven to be a very popular method in many areas of research including bioinformatics and proteomics. It is an ensemble method that uses multiple classification and regression decision trees in its model. The RF builds each decision tree based on a different subset of the data by taking multiple bootstrap samples (sample with replacement) of observations and random samples of variables. For each decision tree in the RF algorithm, the observations that were not used to build the tree are used to validate it.

The performance of a biomarker panel is often assessed using the area under the receiver operator curve (AUC). The AUC is a measure of the degree to which a model can outperform a random classifier. The value of the AUC lies in the range (0–1) with a higher value indicating a better classifier. Notably, when generating an AUC the RF algorithm internally cross-validates the model by using separate training and test datasets from the initial discovery data for each decision tree within the model. The RF output provides cross-validated predicted probabilities of each observation belonging to a response class and these probabilities are then used to generate an AUC value. Hence, all AUC values reported using RF are cross-validated. While the AUC allows us to assess how accurately an algorithm can predict the response variable, it does not establish the identity of the biomarkers in a panel. Instead the members of the panel are usually identified through the use of variable importance measures, which as their name suggests assign importance to individual members of the protein panel. Because it is important to know which members of the biomarker panel contribute to the AUC the RF method has its own default variable importance measure, though this is a somewhat separate part of the classification.

2 Sample sizes and data prefiltering

A widely used approach in proteomic biomarker discovery workflows is to reduce the effect of the "small n large p" problem by prefiltering the variables that come from the initial LC-MS data, so that only the "best" variables are put forward for classification. For example, data are often prefiltered using ANOVA see [17–19]. This "best" filtered subset of data are then often analyzed on the same sample cohort using a classification technique such as RF or those mentioned in Section 1. However, because data from the same sample cohort were used twice: once to choose the filtered subset and again to build a classification model, the results will not give a realistic prediction of the accuracy of the model when it is tested on a different cohort of patients. In our view, this is one of the major limitations of existing proteomic biomarker discovery workflows. This should not be a controversial view as it has been shown in different contexts that prefiltering gives an overly optimistic interpretation of the predictive ability of classification algorithms [5–8]. This happens because the prefiltered variables have the advantage that they already appear to be strongly associated with the response [20, 21]. Here, we use simulated data to illustrate the important considerations of how small sample size and prefiltering lead to model overfitting.

Table 1. Total number of variables, proportion of truly predictive biomarkers, and RF parameters used in the simulation

p	12	112	1 012	10 012
p_{true}/p	0.58	0.0625	0.0069	0.00069
Number of variables sampled by RF	3	10	31	100
Number of trees in RF	500	500	500	500

3 Using simulated data to illustrate the issues of sample size and cross-validation

We consider two criteria as being important for the selection of a good biomarker panel. The first is that the panel accurately predicts the response as estimated by the AUC. The second criterion is that a biomarker panel should contain individual proteins or peptides (variables) that are truly predictive of the response. Clearly, without very large prospective discovery and validation cohorts this latter criterion can only be assessed definitively using simulated data. To test the impact of prefiltering and sample size we created simulated datasets containing authentic predictive variables.

3.1 Simulation model

To create simulated datasets, we adopted the approach of Strobl et al. [22] where data contain seven truly predictive markers (denoted here x_1–x_7 together with five nonpredictive biomarkers (x_8–x_{12}). To mimic real-world behavior of proteins, some of these variables are simulated to be highly correlated. To this initial small dataset, we appended in varying degrees a larger number of random nonpredictive markers. Using this approach, we created simulated data of varying sample sizes, n, (set at 10, 50, and 100) and varying numbers of additional random noise variables (0, 100, 1000, and 10 000). This resulted in a total of 12 datasets that allowed us to determine the impact of data size and prefiltering. Ideally, for each dataset the RF should only choose biomarker panels which, contain the seven authentic markers mentioned above and should not include the nonpredictive markers. We applied the RF algorithm to each dataset and ran it 100 times in order to ensure the reliability of the biomarker panels selected. The total number of variables simulated across datasets (p) as well as the proportion of truly predictive biomarkers (p_{true}/p) can be seen in the top row of Table 1.

3.2 Analysis

Further, for each of the 12 datasets we ran three different scenarios:

(i) RF on the full dataset
(ii) RF after prefiltering by ANOVA with a cutoff p-value of 0.05.
(iii) RF after prefiltering by ANOVA with a Bonferroni correction.

We also performed this analysis using the false discovery rate (FDR). However, in all cases this was found to give near identical results to Bonferroni.

RF requires tuning parameters for the number of variables to sample at each iteration and the number of trees to be used in the forest. Table 1 shows the RF parameters used for each dataset. Table 2 shows the number of variables remaining for each dataset after ANOVA filtering and Bonferroni correction as well as the percentage of truly predictive biomarkers that remained after filtering. To select biomarker panels, the RF model was run using the package random forest and AUC values were calculated using the package ROCR in R version 3.0.1. In each case, the reported performance metrics are obtained as the average cross-validated value over 100 iterations of the RF.

3.3 Predictive and reported accuracy

Figure 1A (top panel) shows the proportion of the seven truly predictive markers correctly found in the top ten most important variables selected by the RF under each scenario. It is clear that when the sample size and the number of variables were small ($n = 10$, $p = 12$, respectively), the selection of truly predictive biomarkers using the full data was very accurate (85.7% average accuracy over the 100 RF iterations). (For the Bonferroni corrected data the p-value threshold (0.05/n) was so strict that it did not accept any variables as significant when the sample size was 10; because of this the first panel of Fig 1B does not include Bonferroni.) However, when the sample size was small and the number of variables was increased ($p > 1000$) this quickly fell away to zero. This figure also shows the effect of increasing sample size on finding truly predictive markers when the number of variables is large. When $p > 1000$ and even with a sample size of 50 only 57.1% of the truly predictive features were found. This shows that finding truly predictive markers in even modest sized proteomic datasets requires that a relatively large number of samples are analyzed.

As mentioned previously, a common way to decrease the number of variables included in the data is to prefilter using methods such as ANOVA and/or correction methods such as Bonferroni. Figure 1A shows that in all but one case, the ANOVA filtered data found the same or fewer of the truly predictive features, with Bonferroni (and equivalently FDR) performing worse in general. The fact that the Bonferroni correction identified fewer of the truly predictive variables is because the number of false positives accepted is reduced (by correcting for multiple testing) and this in turn increases the

Table 2. Total number of variables accepted and percentage of true biomarkers included for full, ANOVA, and Bonferroni datasets for sample sizes 10, 50, and 100

Sample size 10				
Full variables	12 (100%)	112 (100%)	1012 (100%)	10 012 (100%)
ANOVA variables	2 (14.3%)	9 (14.3%)	50 (0%)	478 (0%)
Bonferroni variables	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Sample size 50				
Full variables	12 (100%)	112 (100%)	1012(100%)	10 012(100%)
ANOVA variables	7 (85.7%)	10 (71.4%)	58 (85.7%)	503 (71.4%)
Bonferroni variables	4 (57.1%)	4 (57.1%)	4 (57.1%)	4 (57.1%)
Sample size 100				
Full variables	12 (100%)	112 (100%)	1012 (100%)	10 012 (100%)
ANOVA variables	7 (85.7%)	12 (85.7%)	50 (85.7%)	527 (85.7%)
Bonferroni variables	7 (85.7%)	5 (71.4%)	4 (57.1%)	4 (57.1%)

number of truly predictive biomarkers that are falsely rejected (false negatives).

Figure 1B (bottom panel) shows AUC values obtained for each of the three samples sizes considered (10, 50, 100). It is evident that the ANOVA filtered and Bonferroni corrected data nearly always reported higher accuracy than the analysis performed on the unfiltered full dataset. In fact this accuracy seemed to improve when more noise variables are included in the analysis. For example with a sample size of 50, the reported AUC for ANOVA filtered data was 0.9 for the dataset with 12 variables and increased to 0.99 with 10 012 variables (Fig. 1B) when in fact the actual performance decreased from 85.7 to 57.1% (Fig. 1A).

When the data in Fig. 1A and B are taken together, it is clear that prefiltering data either does not change or reduces the performance of the algorithm, while paradoxically increases the reported AUC values. This deceptive inflation of the AUC for ANOVA filtered and Bonferroni corrected datasets might lead the experimentalist to have undue confidence in the biomarker panel and is likely to lead to wasted time and effort validating biomarker panels that contain considerable random noise. While ANOVA did in many cases

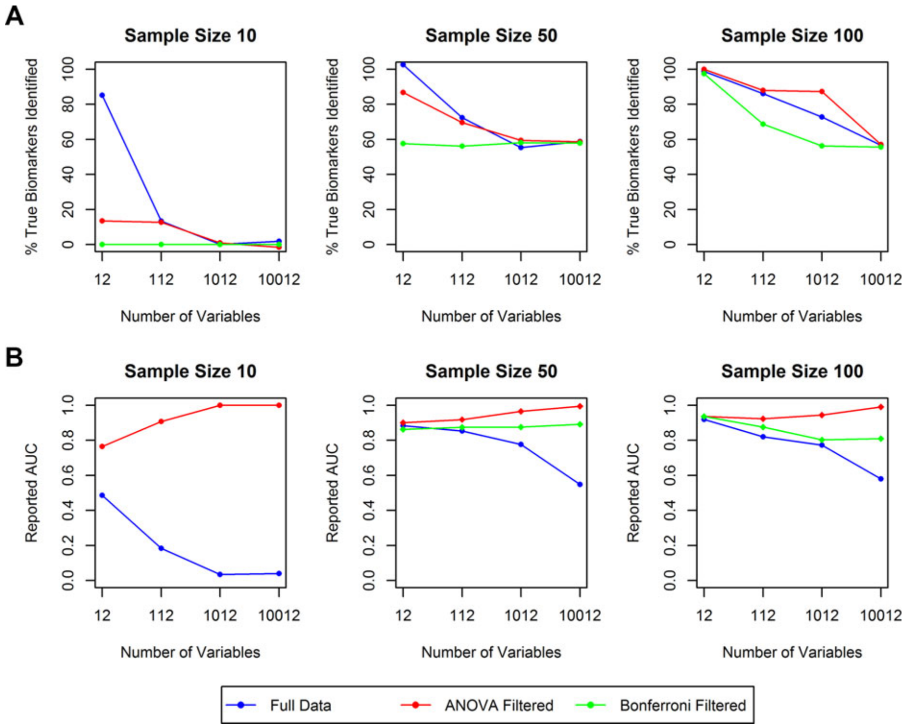

Figure 1. Simulated biomarker datasets. (A; Top panel) Percentage of correctly identified truly discriminating biomarkers. (B; Bottom panel) AUC values for the full and ANOVA and Bonferroni filtered datasets.

identify the same number of true biomarkers as were found through analysis of the full dataset; the consequences of the overinflated reported accuracy when using ANOVA or Bonferroni here negate the benefits of simplifying the data. Also, it should be noted that, ANOVA and hence Bonferroni and FDR would be expected to perform well for this simulation as the variables are linearly related. If the data were nonlinear, as might reasonably be expected, we would anticipate an even worse performance from these methods.

4 Concluding remarks

Through the use of simulated datasets, we have illustrated that more robust identification of candidate biomarkers will result from the use of larger sample sizes in proteomic biomarker discovery experiments. This is a point that is often noted but it seems rarely acted upon. It is our view that this contributes significantly to the poor record of proteomics for delivering clinically validated biomarkers, because authentic candidates are not selected from discovery experiments. Importantly, we have also shown that the commonly used approach of prefiltering of the initial discovery data by ANOVA and correction methods like Bonferroni and FDR rarely improve the accuracy of biomarker selection. Hence, we propose that greater time and attention to appropriate statistical considerations are made earlier in the biomarker discovery and development process. Further, we suggest that only when the data are not prefiltered, can the quality of a biomarker panel be accurately judged through the AUC. If prefiltering is performed, it should be done outside of the model cross-validation process or by having a separate external cross-validation for filtering and an internal cross-validation for accessing model performance. In other words, ANOVA filtering should be performed on a completely separate cohort to that used to assess the performance of the chosen biomarker panel in order to avoid inflation of the performance metrics.

Sample size is also a major consideration. We suggest that LC-MS clinical biomarker discovery experiments are undertaken on the maximum number of samples possible, preferably at least 50 in order to (i) minimize the effects of overfitting and (ii) improve the quality of performance metrics. Although it may be argued that it is not feasible to run such relatively large sample sizes, we suggest that investing at this stage in the process will ultimately save significant time and effort that might otherwise be wasted in validating a nonpredictive biomarker panel. It is notable that this validation often requires the time-consuming assembly of highly valuable and precious patient cohorts.

This work was supported by the Irish Research Council. Funding for mass spectrometry instrumentation was supported by Science Foundation Ireland. Protein biomarker discovery projects undertaken in the Pennington laboratory are supported by the Health Research Board (grant number HRA_POR/2011/125), the Irish Cancer Society (grant number PCI11WAT), Movember GAP and St. Luke's Institute for Cancer Research. The UCD Conway Institute and Proteome Research Centre are funded by the Program for Research in Third Level Institutions as administered by the Higher Education Authority of Ireland. We are grateful to the reviewers whose comments and criticisms enabled us to distil the key elements from our extensive data simulation studies.

The authors have declared no conflict of interest.

5 References

[1] Diamandis, E. P., The failure of protein cancer biomarkers to reach the clinic: why, and what can be done to address the problem? *BMC Med.* 2012, *10*, 87.

[2] Oon, S., Pennington, S., Fitzpatrick, J., Watson, R. W. G., Biomarker research in prostate cancer-towards utility, not futility. *Nat. Rev. Urol.* 2011, *8*, 131–138.

[3] Rifai, N., Gillette, M. A., Carr, S. A., Protein biomarker discovery and validation: the long and uncertain path to clinical utility. 2006, *24*, 971–983.

[4] Kaiser, J., Clinical medicine. Biomarker tests need closer scrutiny, IOM concludes. *Science* 2012, *335*, 1554.

[5] Hilario, M., Kalousis, A., Pellegrini, C., Müller, M., Processing and classification of protein mass spectra. *Mass Spectrom. Rev.* 2006, *25*, 409–449.

[6] Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S. et al., Sample classification from protein mass spectrometry, by "peak probability contrasts". *Bioinformatics* 2004, *20*, 3034–3044.

[7] Robin, X., Turck, N., Hainard, A., Lisacek, F., Sanchez, J., Bioinformatics for protein biomarker panel classification: what is needed to bring biomarker panels into in vitro diagnostics? *Expert Rev. Proteomics* 2009, *6*, 675–679.

[8] Ambroise, C., McLachlan, G. J., Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. U S A* 2002, *99*, 6562–6566.

[9] Wright, M. E., Han, D. K., Aebersold, R., Mass spectrometry-based expression profiling of clinical prostate cancer. *Mol. Cell. Proteomics* 2005, *4*, 545–554.

[10] Surinova, S., Schiess, R., Hüttenhain, R., Cerciello, F. et al., On the development of plasma protein biomarkers. *J. Proteome Res.* 2011, *10*, 5–16.

[11] Podwojski, K., Eisenacher, M., Kohl, M., Turewicz, M. et al., Peek a peak: a glance at statistics for quantitative label-free proteomics. *Proteomics* 2010, *7*, 249–261.

[12] Sampson, D. L., Parker, T. J., Upton, Z., Hurst, C. P., A comparison of methods for classifying clinical samples based on proteomics data: a case study for statistical and machine learning approaches. *PLoS One* 2011, *6*, e24973.

[13] Wu, B., Abbott, T., Fishman, D., McMurray, W. et al., Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 2003, *19*, 1636–1643.

[14] Chen, B., Sheridan, R., Hornak, V., Voigt, J. H., Comparison of random forest and Pipeline Pilot Naïve Bayes in

prospective QSAR predictions. *J. Chem. Inf. Model.* 2012, *52*, 792–803.

[15] Caruana, R., Karampatziakis, N., Yessenalina, A., *An empirical evaluation of supervised learning in high dimensions.* Proceedings of the 25th International Conference on Machine Learning (ICML '08), Helsinki. 2008, pp. 96–103.

[16] Breiman, L., Random forests. *Mach. Learn.* 2001, *45*, 5–32.

[17] Böhm, D., Keller, K., Wehrwein, N., Lebrecht, A. et al., Serum proteome profiling of primary breast cancer indicates a specific biomarker profile. *Oncol. Rep.* 2011, 1051–1056.

[18] Long, L., Li, R., Li, Y., Hu, C., Li, Z., Pattern-based diagnosis and screening of differentially expressed serum proteins for rheumatoid arthritis by proteomic fingerprinting. *Rheumatol. Int.* 2011, *31*, 1069–1074.

[19] Chen, F., Xue, J., Zhou, L., Wu, S., Chen, Z., Identification of serum biomarkers of hepatocarcinoma through liquid chromatography/mass spectrometry-based metabonomic method. *Anal. Bioanal. Chem.* 2011, *401*, 1899–904.

[20] Lausser, L., Müssel, C., Maucher, M., Kestler, H. a., Measuring and visualizing the stability of biomarker selection techniques. *Comput. Stat.* 2011, *28*, 51–65.

[21] Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K., Boulesteix, A.-L., Over-optimism in bioinformatics: an illustration. *Bioinformatics* 2010, *26*, 1990–1998.

[22] Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., Conditional variable importance for random forests. *BMC Bioinformatics* 2008, *9*, 307.