Published in final edited form as:
*J Proteome Res*. 2009 February ; 8(2): 1030–1036. doi:10.1021/pr800832s.

# Prion Disease Diagnosis by Proteomic Profiling

**Allen Herbst**[1,†], **Sean McIlwain**[2,†], **Joshua J. Schmidt**[3,†], **Judd M. Aiken**[1], **C. David Page**[2,4], and **Lingjun Li**[3,5,*]

[1]*Department of Comparative Biosciences, University of Wisconsin- Madison*

[2]*Department of Computer Sciences, University of Wisconsin-Madison*

[3]*School of Pharmacy, Division of Pharmaceutical Sciences, University of Wisconsin-Madison*

[4]*Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison*

[5]*Department of Chemistry, University of Wisconsin-Madison*

## Abstract

Definitive prion disease diagnosis is currently limited to post-mortem assay for the presence of the disease-associated proteinase K-resistant prion protein. Using cerebrospinal fluid (CSF) from prion-infected hamsters, matrix-assisted laser desorption/ionization Fourier transform mass spectrometry (MALDI-FTMS), and support vector machines (SVM), we have identified peptide profiles characteristic of disease state. Using ten-fold leave-one-out cross-validation, we report a predictive accuracy of 72% with a true positive rate of 73% and a false positive rate of 27% demonstrating the suitability of using proteomic profiling and CSF for the development of multiple marker diagnostics of prion disease.

### Keywords

Biomarkers; Prion Disease; Transmissible Spongiform Encephalopathy; MALDI-FTMS; Machine Learning Algorithms; Support Vector Machines; Proteomics

## Introduction

Prion diseases (Transmissible Spongiform Encephalopathies (TSEs)) are a unique family of fatal neurodegenerative diseases that affect mammals. They include scrapie in sheep and goats, bovine spongiform encephalopathy (mad cow disease or BSE), transmissible mink encephalopathy, and chronic wasting disease (CWD) in elk and deer. Human forms of the disease include: genetic disease, Gerstmann-Sträussler-Scheinker syndrome and fatal familial insomnia; sporadic disease, Creutzfeldt-Jakob disease (CJD), and infectious disease, variant CJD (vCJD), caused by the consumption of BSE infected cattle, and kuru, linked to the practice of ritualistic cannibalism in Papua New Guinea. Clinical features of TSEs vary with host species but, generally, affected animals display pruritus, ataxia, and ultimately, death [1] following an extended asymptomatic incubation period of months to decades [2] during which infectious agent can replicate to very high titers (>1×$10^8$ infectious units). Histopathological features of TSEs are characterized by spongiform degeneration, reactive astrocytosis and the accumulation of amyloid in the central nervous system.

*Correspondence to: Professor Lingjun Li, School of Pharmacy, University of Wisconsin, 777 Highland Avenue, Madison, Wisconsin 53705−2222 USA; Phone: 608−265−8491; Fax: 608−262−5345; Email: lli@pharmacy.wisc.edu.

†each of these authors contributed equally to this work

TSEs are characterized at the molecular level by the accumulation of an abnormal aggregated isoform, $PrP^{scrapie}$ ($PrP^{Sc}$), of the prion protein, $PrP^{cellular}$ ($PrP^{C}$), in affected animals. $PrP^{C}$ is a 33–35 kDa protein encoded by a single copy gene [3, 4]. During the course of a TSE infection, $PrP^{C}$ undergoes a post-translational conformational conversion to a disease-specific isoform ($PrP^{Sc}$) that has increased β-sheet content, resistance to proteinase K digestion and detergent insolubility.

The outbreaks of BSE, subsequently variant CJD, and emerging prion diseases such as CWD have prompted the need for rapid and reliable ante-mortem screening methods that allow definitive TSE diagnosis at both symptomatic and pre-symptomatic stages. Current diagnostic tests are typically performed post-mortem and use immunohistochemistry or enzyme-linked immunosorbent assay (ELISA) to detect abnormal prion protein from diseased brains or lymphoid tissues. Although the antibody-based tests are precise and accurate, an ante-mortem method of diagnosis that could be performed using a body fluid would be of great utility. A reliable pre-mortem diagnostic test for animals infected with prion diseases is lacking due to the low level of infectious prions present in bodily fluids suitable for diagnostic analysis (blood/serum, cerebrospinal fluid, or urine) coupled with the analytical difficulty associated with detecting the disease specific protein conformation, $PrP^{Sc}$, against a background of $PrP^{C}$. The "gold standard" for prion diagnostics is immunohistochemistry utilizing anti-prion protein antibodies on the obex region of the brain [5]. Drawbacks to immunohistochemistry include the low throughput of samples and the necessity for post-mortem analysis. Other antibody-based diagnostics, such as the Prionics or Bio-Rad tests, utilize a Western Blot/ELISA approach that take advantage of the protease resistance of the abnormal form of the prion protein. Despite the good specificity and sensitivity of these post-mortem tests, animals infected with prion disease cannot be diagnosed until late in the pre-clinical period when sufficient abnormal PrP has accumulated. The ultimate goal should be the development of pre-clinical diagnostics that can facilitate detection of disease before potentially contaminated food or blood products enter the market. Furthermore, the ability to detect and quantify disease progression allows for the evaluation of therapeutic strategies without requiring death as an end-point.

Surrogate markers of prion disease have been identified in patients presenting clinical signs of CJD. Among these are the characteristic electroencephalogram pattern observed in CJD cases, the presence of central nervous system-specific protein markers such as 14–3–3, tau, apolipoprotein E, cystatin C [6-10] and neuron-specific enolase [8, 11, 12]. Currently, high levels of 14–3–3 proteins present in the cerebrospinal fluid (CSF) serve as the clinical criteria for the diagnostic upgrade of suspect CJD to probable CJD, with confirmed CJD only being diagnosed post-mortem. Despite being secondary to post-mortem immunohistochemistry for diagnosis due to low disease specificity, these markers indicate a robust biological response to prion infection during clinical disease.

Mass spectrometry is an increasingly popular tool for protein biomarker discovery due to its high sensitivity, speed, chemical specificity and capability for complex mixture analysis. At the heart of this rapidly evolving research area is the capability to characterize an ensemble of proteins expressed in a tissue or secreted into body fluids. These new capabilities became possible due to major innovations in ionization methodologies (matrix-assisted laser desorption/ionization [13, 14] and electrospray ionization [15]). With these advancements, peptides or proteins in condensed phase could be converted to intact gas-phase ions for mass measurement. Coupled with the development of various mass analyzers [16], tandem mass spectrometry (MS/MS) experiments can be conducted to produce amino-acid sequence specific fragment ions to allow identification of peptides/proteins of interest.

One of the widely used proteomic platforms, surface-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) MS, was among the first MS technologies to monitor protein expression profiles in diseased tissues and body fluids [17-19]. While highly sensitive and extensively used in diagnostics, the SELDI-TOF MS platform suffers from several limitations, including limited dynamic range and poor resolution and mass accuracy. Recognizing these limitations, there is a great deal of effort directed at exploring alternative MS platforms for biomarker discovery[20-25]. Among these developments, Fourier transform mass spectrometry (FTMS) is emerging as an attractive platform due to the high resolving power, mass measurement accuracy, multi-stage MS/MS capabilities and large dynamic range [26-31]. When coupled with 1D or 2D separations, FTMS shows significant promise to address the sample complexity encountered in biomarker discovery research [32-36].

There are many issues to consider when analyzing proteomic mass spectrometric data for classification. For example, analyte suppression from high abundance species could obscure the detection of low-level analyte(s) of interest in a highly complex mixture. Detector saturation reduces the predictive power of the peak intensities of the sample's mass spectrum. There are mass-to-charge (m/z) shifts of common peaks between spectra and within spectra; there are noise peaks that are not indicative of the underlying biology. These noise peaks are caused by chemical and electronic noise during the mass spectrometry measurement. Also, there are redundant features due to isotopic distributions, various adducts, multiple charge states, and peptide fragments occurring from proteolysis.

The application of machine learning to problems in computational biology has been widely studied in a number of different areas. Applications to genetic microarray analysis [37], protein structure, and folding prediction [38, 39], and biomedical text analysis [40, 41] are example problems where machine learning has made contributions. Often this involves applying a statistical modeling algorithm to a set of biological data. Machine learning provides a series of techniques and modeling algorithms taken from the artificial intelligence domain and applied to the data. These statistical models attempt to “learn” the underlying concept rather than fit a provided model. These modeling techniques take as input a set of examples (data set), each described by a feature vector. For mass spectrometry data, the feature vector is the series of peaks that define the underlying spectra. Using this feature vector, the learned model determines an output value, which is indicative of the concept that is desired. Example tasks include classification and regression. Classification models try to learn a discrete output from the supplied feature vector. A concept is generally of binary value (Yes or No), but can also be more than two-class. Examples of classification models include decision trees, Bayesian networks, and Support Vector machines [42-44]. Some classifiers can also associate a probability to the output values. Thresholding upon this probability allows for the adjustment of the classifier's performance metrics. Support vector machines (SVMs) classify examples by plotting the example features into a high dimensional space and finding a hyper-plane that best separates the positive from the negative examples. The hyper-plane parameters are optimized to minimize the number of classification errors of the training set.

Here we describe a multidisciplinary approach to identify ante-mortem markers of prion disease. Our strategy combines MALDI-FTMS for accurate mass fingerprinting of peptide mixtures derived by tryptic digestion and machine learning for classification of mass spectral features collected from CSF samples obtained from preclinical prion-infected and uninfected animals.

## Materials and Methods

### Sample Preparation

Weanling Syrian golden hamsters were inoculated perorally with 50μL of a 10% brain homogenate of 263K prion agent for 5 days. Mock-inoculated uninfected animals served as controls. CSF was drawn by lumbar puncture at 18 weeks after inoculation from both infected (n=21) and uninfected animals (n=22). CSF was collected from infected and uninfected animals on the same day, alternating between infected and uninfected, to minimize temporal bias. The typical volume of CSF collected was approximately 10μL. Following collection, samples were digested with sequencing-grade modified trypsin (Promega Co., Madison, WI) in 50% acetonitrile/ammonium bicarbonate buffer (15mM) overnight [45]. The resulting peptide mixtures were co-crystallized with an equal volume of DHB matrix, a 1:1 mixture of aqueous 2,5-dihydroxybenzoic acid (DHB) and methanol on a stainless steel sample target. Each sample was spotted three times and each spot was analyzed three times using MALDI-FTMS resulting in nine spectra being produced for each sample, giving a total of 387 spectra for both infected and control animals. Like CSF, spectra were collected to avoid introducing temporal bias into the data set.

## MALDI-FTMS

Spectra were obtained using an IonSpec 7.0T matrix-assisted laser desorption/ionization – Fourier transform mass spectrometer, or MALDI-FTMS. Samples and matrix were irradiated using a nitrogen laser emitting light at 337 nm. The DHB matrix absorbed this light and caused the sample to ionize and desorb into the gas phase. A quadrupole ion guide carried analytes into an ion cyclotron resonance cell where an RF scan was applied to excite the ions at an amplitude of 150 V base to peak. The filament and quadrupole trapping plates were initialized to 15 V, and subsequently decreased to 1V in 6.5 to 7.0 seconds to reduce baseline distortion of peaks. Detection was performed in broadband mode from *m/z* 108.0 to 4500.0.

## Data Processing and Analysis

Raw peaklists derived from the Ion Spec peak picking software were collected from prion-infected and uninfected hamster CSF samples and processed as follows: spectra were aligned by using internal reference peaks at *m/z* 673.1, 1274.6, 2060.0, 2126.0, and 2326.0, peaks below *m/z* 500.0 were removed, and, the spectra were de-isotoped using a trained model based upon hand-annotated spectra. The de-isotoping model was constructed using dynamic programming and is an extension of the algorithm from McIlwain et al. 2007 [46]. The algorithm accounts for the possibility that isotopic distributions of peaks derived from different proteins/peptides may interleave with each other on the mass-to-charge axis. The parameters for the de-isotoping algorithm are: valid isotope cutoff of 150 ppm, the maximum charge state is 1 given the ionization technique, minimum peaks for a valid isotope of two, and training the algorithm parameters on a grid of 0−1 for noise penalty, 0−0.9 for noise threshold, and 0−1.0 for the overlap penalty along with a search length from 5−12 maximizing the F1 score of the Mono-Isotopic Fine [47] and the classifier employed is naïve Bayes. The valid isotope cutoff is not used to identify isotopic distributions, but rather to decrease processing time of the algorithm. Peaks that lack an isotopic distribution were removed (as these cannot be biologically derived and were likely the result of electrical noise). Each annotated isotopic distribution was collapsed to the most intense peak. De-isotoped spectra were converted into a feature-based data set by cluster aligning isotopic distributions across spectra centered on the most intense peak of each distribution and establishing bins. Feature values (peptide peaks) to be used in classification were selected based on their presence in at least three out of the nine spectra taken from each animal. A minimum threshold of three peaks per bin was selected to ensure repeatable peaks were included in the classification model. Higher thresholds were inappropriate because they caused the classification model to be constructed using abundant features. To determine the number of features to use, we tuned the algorithm with an inner loop of cross-validation (10-fold) using 1, 5, 10, 50, 100, 500, 1000, 1500, 2000, 2500, and all features. We then trained and classified upon the selected features using linear support vector machines (SVM). The classifier that we used in this work is the linear SVM provided by the weka java package [48].

## Results and Discussion

We combined mass spectrometry and machine learning algorithms to identify differentially expressed peptides and peptide profiles from hamster CSF obtained from animals with preclinical prion disease. We used CSF from the hamster model of prion disease, 18 weeks post-inoculation because disease progression is well-defined in this model; there is no clinical manifestation of disease at this time; and CSF can be collected in sufficient volumes. At 18 weeks post-inoculation, 82% of the incubation period has passed. Infectivity has reached the maximum titer in the brain and astrocytosis can be observed using anti-glial fibrillary acidic protein antibodies. Ependymal cells in the choroid plexus of the brain ventricles produce CSF. CSF is the only fluid in direct contact with the brain and, thus, is a potential source of biomarkers for CNS disorders. CSF has been the subject of several MS-based proteomic studies in neurodegenerative diseases during recent years [49, 50].

On average, we obtained very good resolution of peptide peaks with minimal background and chemical noise. As expected, the gross protein profiles of both infected (Figure 1A) and uninfected (Figure 1B) samples produce very similar sets of peaks due to the presence of non-differentially regulated proteins. These peptides tend to be non-variable as a result of their high abundances across all samples. At closer inspection, however, differentially regulated peptides can be observed among the less abundant peaks (Figure 1). Less abundant peaks were far more variable amongst the replicate spectra from across and within each sample spot, consistent with the stochastic chance of those peptides being ionized and detected above background. We used peak lists derived from the IonSpec peak picking software as the basis for our analysis. The peptides that constitute these lists were de-isotoped to remove redundant features as well as to identify those features that are derived from biological sources. The de-isotoping algorithm that processes these spectra was trained on a set of ten annotated peak lists generated by hand. Collapsing to the most intense peak ensures that the clustering algorithm properly aligns the resultant classifying features. To evaluate the performance of our de-isotoping algorithm, we generated mono-isotopic [46] and mono-isotopic fine scores [47]. The resulting Mono-Isotopic and Mono-Isotopic Fine scores are averaged from leave-one-out (10-fold) cross-validation of the 10 expert-annotated peak lists. Our de-isotoping scores were mono-isotopic (F1 − 78 +/− 10%, Precision − 79 +/− 16%, Recall − 81 +/− 7%) and mono-isotopic fine (F1 − 80 +/− 10%, Precision − 80 +/− 15%, Recall − 84 +/− 6%).

To transform the pre-processed peak lists into feature lists, individual bins were created across spectra and the presence or absence of a peak was determined in each bin from each spectrum. Those bins that contained a minimum of 3 peaks from each sample were defined as positive and individual bins were ranked based upon the information gain metric. The effect of this is to exclude those peaks from the feature data and subsequent classification model that do not possess minimum information gain scores. Information gain is a concept used to estimate the reduction in complexity of a problem given a particular suggestion [42]. In our case, the problem is detecting disease state and the particular suggestion is the presence or absence of a given peak. As can be seen in Figure 2, the information gain scores obtained as a result of randomized class labels is lower suggesting that the best classification of spectra is based on disease state. From ranked individual bins (Figure 2), feature sets are created consisting of the top features from both the infected and uninfected classifications or artificial classifications based upon randomizations of either animals or spectra. While the individual detection of a peptide may be highly correlated with disease, more sensitive and specific classification can occur by using the information in multiple peaks. To identify a reasonable number of features to include in the classification model, we examined the relationship between overall accuracy and the number of features. Figure 3 demonstrates the increase in accuracy obtained by using multiple features to establish disease state. As the number of features increases from 0 to 100, the overall accuracy improves demonstrating the utility of using multiple features for disease classification. Incorporating additional features, however, does not improve accuracy substantially. As a result, we tuned the number features based upon accuracy for each inner fold of the cross-validation. Based on the performance evaluation and typical number of peptide peaks detected in the hamster CSF mass spectra, a range of 2688−2807 peptides were used for clustering and machine learning.

After identifying useful features, we used leave-one-out cross-validation to estimate the predictive accuracy of our diagnostic method based on linear SVM and naïve Bayes machine learning algorithms. Classification models are trained on infected and uninfected samples while one sample is held out for testing. The entire process of feature selection and machine learning was repeated on every fold of the cross-validation without looking at the held-out sample for that fold. The resulting model is then tested on the held-out sample. In other words, after the model is built from the n-1 set of samples, called the training set, we measure the performance of the model's predictions on the held out sample, called the test set. We repeat this process n times, holding out a different sample to build a confusion matrix (Figure 4) of the prediction results.

Using this leave-one-out cross-validation, the model's specificity and sensitivity are tested. We found a predictive accuracy of 72% with a true positive rate (sensitivity) of 73% and a false positive rate of 27% (specificity 73%). Another way of measuring performance is through the use of receiver operator characteristic (ROC) curves (Figure 5). Using a probabilistic classifier, the probability of an example being positive is given instead of a yes/no classification. By setting a threshold on this probability, a series of points for the True Positive Rate (Recall) and False Positive Rate is obtained. Plotting these generates an ROC curve. This curve helps describe the tradeoff between the True Positive Rate and the False Positive Rate of the method.

To formally test our classification model, we employed permutations of the labels in the data set, generating one thousand random groups of spectra with artificially assigned labels of infected or uninfected. By performing this permutation test (Figure 6), we can estimate the probability that disease state is the best interpretation of the full data set, rigorously evaluating whether our model is overly optimistic in its evaluation of disease class. Classifying the randomized data sets using the same methods as in the true infected/uninfected case produces, roughly, a normal distribution for accuracy scores. The probability of obtaining an accuracy of 72% is less than 0.007 suggesting that grouping the data set into infected and uninfected animals is the best arrangement, consistent with the true biological test (Figures 5, 6).

Despite a low probability to obtain an accuracy of 72%, numerous biological, analytical and informatic challenges remain. One of the most important is the identification of specific biomarkers and the ascertainment of their role in prion disease pathogenesis. The hamster model, while very useful for studies of prion disease infectivity, is limited by the paucity of genomic and proteomic bioinformatic knowledge, limiting the identification of specific peptides and proteins based upon their masses. Furthermore, the volume of CSF obtainable from hamsters prevents fractionation schemes that could reduce protein complexity which is critical for the detection of lower abundance proteins. Clearly, however, a disease-specific protein signature exists in the CSF, indicating the potential of this approach to prion diagnostics. Testing these profiles against other models of neurodegeneration would provide a measure of prion-disease specificity. Future technological enhancements in sample preparation, protein mass detection, and bioinformatic analysis will yield improvements in accuracy, sensitivity, specificity, and confidence in the validity of a multiple protein approach to disease detection and, ultimately, monitoring of disease progression.

## Conclusions

We have demonstrated that biomarkers of prion infection can be identified from the CSF of hamsters at a pre-clinical time-point, suggesting the utility of our approach to ante-mortem diagnostics. Our use of multiple markers of disease state has enhanced the sensitivity and specificity of our classification yielding higher accuracies in determining disease state. The disease classification model is largely superior to classifications arising from randomization of the data set suggesting that the best interpretation is that some animals have preclinical prion disease and others do not.

## Acknowledgements

We are grateful to the University of Wisconsin School of Pharmacy Analytical Instrumentation Center for access to the MALDI-FTMS instrument. This work was supported in part by the National Prion Research Program, DAMD17 −03−1−0291, the National Institutes of Health, AI0272588, and the Wisconsin Alumni Research Foundation at the University of Wisconsin-Madison. L.L. acknowledges an Alfred P. Sloan Research Fellowship.

## References

1. Detwiler LA. Revue Scientifique et Technique 1992;11:491–537. [PubMed: 1617202]
2. Hunter GD. Journal of Infectious Diseases 1972;125:427–440. [PubMed: 4622954]
3. Chesebro B, Race R, Wehrly K, Nishio J, Bloom M, Lechner D, Bergstrom S, Robbins K, Mayer L, Keith JM. Nature 1985;315:331–333. [PubMed: 3923361]
4. Oesch B, Westaway D, Walchli M, McKinley MP, Kent SB, Aebersold R, Barry RA, Tempst P, Teplow DB, Hood LE. Cell 1985;40:735–746. [PubMed: 2859120]
5. Schaller O, Fatzer R, Stack M, Clark J, Cooley W, Biffiger K, Egli S, Doherr M, Vandevelde M, Heim D, Oesch B, Moser M. Acta Neuropathologica 1999;98:437–443. [PubMed: 10541864]
6. Van Everbroeck B, Boons J, Cras P. Clin Neurol Neurosurg 2005;107:355–360. [PubMed: 16023527] Epub 2005 Jan 2012
7. Parveen I, Moorby J, Allison G, Jackman R. Vet Res 2005;36:665–683. [PubMed: 16120244]
8. Otto M, Wiltfang J, Cepek L, Neumann M, Mollenhauer B, Steinacker P, Ciesielczyk B, Schulz-Schaeffer W, Kretzschmar HA, Poser S. Neurology 2002;58:192–197. [PubMed: 11805244]
9. Choe LH, Green A, Knight RS, Thompson EJ, Lee KH. Electrophoresis 2002;23:2242–2246. [PubMed: 12210228]
10. Sanchez JC, Guillaume E, Lescuyer P, Allard L, Carrette O, Scherl A, Burgess J, Corthals GL, Burkhard PR, Hochstrasser DF. Proteomics 2004;4:2229–2233. [PubMed: 15274116]
11. Burkhard PR, Sanchez JC, Landis T, Hochstrasser DF. Neurology 2001;56:1528–1533. [PubMed: 11402110]
12. Zerr I, Bodemer M, Otto M, Poser S, Windl O, Kretzschmar HA, Gefeller O, Weber T. Lancet 1996;348:846–849. [PubMed: 8826809]
13. Karas M, Hillenkamp F. Anal Chem 1988;60:2299–2301. [PubMed: 3239801]
14. Hillenkamp F, Karas M, Beavis RC, Chait BT. Anal Chem 1991;63:1193A–1203A. [PubMed: 1897719]
15. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Science 1989;246:64–71. [PubMed: 2675315]
16. Aebersold R, Mann M. Nature 2003;422:198–207. [PubMed: 12634793]
17. Issaq HJ, Veenstra TD, Conrads TP, Felschow D. Biochem Biophys Res Commun 2002;292:587–592. [PubMed: 11922607]
18. Petricoin EF, Liotta LA. Curr Opin Biotechnol 2004;15:24–30. [PubMed: 15102462]
19. Merchant M, Weinberger SR. Electrophoresis 2000;21:1164–1177. [PubMed: 10786889]
20. Hamler RL, Zhu K, Buchanan NS, Kreunin P, Kachman MT, Miller FR, Lubman DM. Proteomics 2004;4:562–577. [PubMed: 14997480]
21. Rodland KD. Dis Markers 2004;20:129–130. [PubMed: 15502244]

22. Simoneit BR. Mass Spectrom Rev 2005;24:719–765. [PubMed: 15534872]

23. Kolch W, Neususs C, Pelzing M, Mischak H. Mass Spectrom Rev 2005;24:959–977. [PubMed: 15747373]

24. Koomen JM, Li D, Xiao LC, Liu TC, Coombes KR, Abbruzzese J, Kobayashi R. J Proteome Res 2005;4:972–981. [PubMed: 15952745]

25. Reyzer ML, Caprioli RM. J Proteome Res 2005;4:1138–1142. [PubMed: 16083264]

26. Gross ML, Rempel DL. Science 1984;226:261–268. [PubMed: 6385250]

27. Li Y, McIver RT Jr. Hunter RL. Anal Chem 1994;66:2077–2083. [PubMed: 8067524]

28. Marshall AG, Hendrickson CL, Jackson GS. Mass Spectrom Rev 1998;17:1–35. [PubMed: 9768511]

29. McLafferty FW, Kelleher NL, Begley TP, Fridriksson EK, Zubarev RA, Horn DM. Curr Opin Chem Biol 1998;2:571–578. [PubMed: 9818181]

30. Castoro JAW, C. L. Trends Anal Chem 1994;13:229–233.

31. Tseng K, Xie Y, Seeley J, Hedrick JL, Lebrilla CB. Glycoconj J 2001;18:309–320. [PubMed: 11788799]

32. Bergquist J, Palmblad M, Wetterhall M, Hakansson P, Markides KE. Mass Spectrom Rev 2002;21:2–15. [PubMed: 12210611]

33. Ramstrom M, Ivonin I, Johansson A, Askmark H, Markides KE, Zubarev R, Hakansson P, Aquilonius SM, Bergquist J. Proteomics 2004;4:4010–4018. [PubMed: 15540204]

34. Bergen HR 3rd, Klug MG, Bolander ME, Muddiman DC. Rapid Commun Mass Spectrom 2004;18:1001–1002. [PubMed: 15116428]

35. Chalmers MJ, Mackay CL, Hendrickson CL, Wittke S, Walden M, Mischak H, Fliser D, Just I, Marshall AG. Anal Chem 2005;77:7163–7171. [PubMed: 16285662]

36. Hawkridge AM, Heublein DM, Bergen HR 3rd, Cataliotti A, Burnett JC Jr. Muddiman DC. Proc Natl Acad Sci U S A 2005;102:17442–17447. [PubMed: 16293687]Epub 12005 Nov 17417

37. Berrar DP, Downes CS, Dubitzky W. Pac Symp Biocomput 2003:5–16. [PubMed: 12603013]

38. Kim S. Bioinformatics 2004;20:40–44. [PubMed: 14693806]

39. Chivian D, Kim DE, Malmstrom L, Schonbrun J, Rohl CA, Baker D. Proteins 2005;61:157–166. [PubMed: 16187358]

40. Cohen AM, Hersh WR. Brief Bioinform 2005;6:57–71. [PubMed: 15826357]

41. Wang H, Huang M, Ding S, Zhu X. BMC Bioinformatics 2008;9:S4.

42. Mitchell, TM. Machine Learning. McGraw-Hill; New York: 1997.

43. Schölkopf, B.; Burges, CJC.; Smola, AJ. Advances in kernel methods : support vector learning. MIT Press; Cambridge, Mass.: 1999.

44. Cristianini, N.; Shawe-Taylor, J. An introduction to support vector machines : and other kernel-based learning methods. Cambridge University Press; Cambridge ; New York: 2000.

45. Strader MB, Tabb DL, Hervey WJ, Pan C, Hurst GB. Anal Chem 2006;78:125–134. [PubMed: 16383319]

46. McIlwain S, Page D, Huttlin EL, Sussman MR. Bioinformatics 2007;23:i328–336. [PubMed: 17646314]

47. McIlwain S, Page D, Huttlin EL, Sussman MR. Bioinformatics 2008;24:i339–347. [PubMed: 18586733]

48. Holmes G, Donkin A, Witten IH. Proceedings Second Australian and New Zealand Conference on Intelligent Information Systems 1994:357–361.

49. Yuan X, Desiderio DM. Proteomics 2005;5:541–550. [PubMed: 15627968]

50. Yuan X, Desiderio DM. J Mass Spectrom 2005;40:176–181. [PubMed: 15706611]

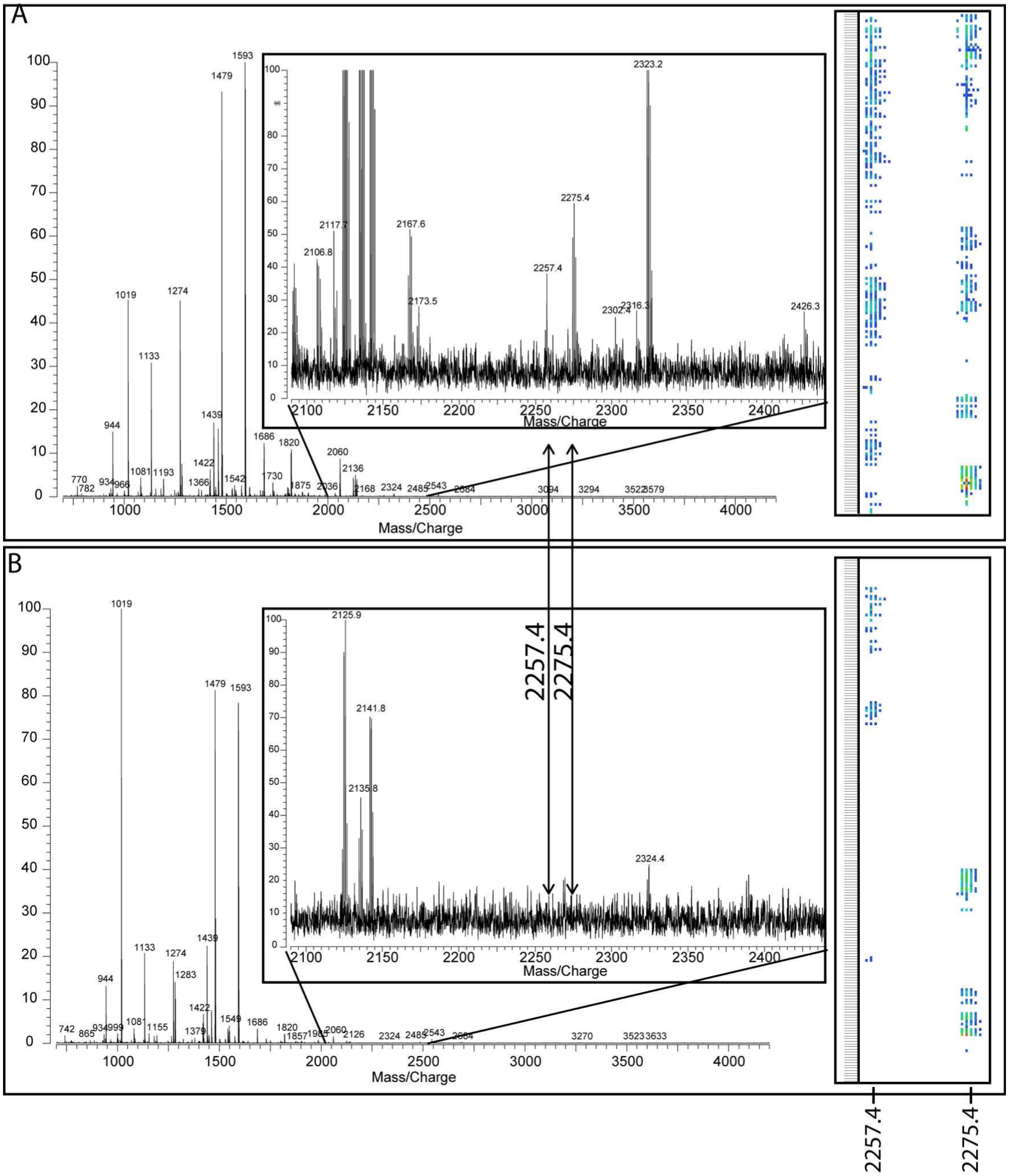


**Figure 1.**

Identification of putative biomarkers of prion disease using MALDI FTMS profiling of tryptic peptides of CSF and heat map feature extraction. Representative mass spectra of hamster CSF, infected (A) and uninfected control (B), collected 18 weeks post-inoculation. The blow up boxes are zoomed on a peptide (*m/z* =2275.4) that is present in infected hamsters. The peak at *m/z* 2257.4 is the dehydrated peptide. The panel on the right shows the distribution of this biomarker among infected and control animals. Each dash represents one of nine individual spectra collected in replicates from 21 infected or 22 uninfected hamsters. Colored pixels to the right indicate detection of a peak at the given mass. Color heat is indicative of the relative intensity of the peak within its spectrum. The multiplicity of pixels is indicative of the peptide's isotopic distribution. The peptide is present in 15% of uninfected animals and 60% of infected animals. Although this feature can be differentiated in these two spectra, it is not consistently present in the infected or absent in the uninfected and is therefore cannot be used alone as an indicative feature, thus requiring the consideration of multiple features as for accurate diagnosis of disease state.

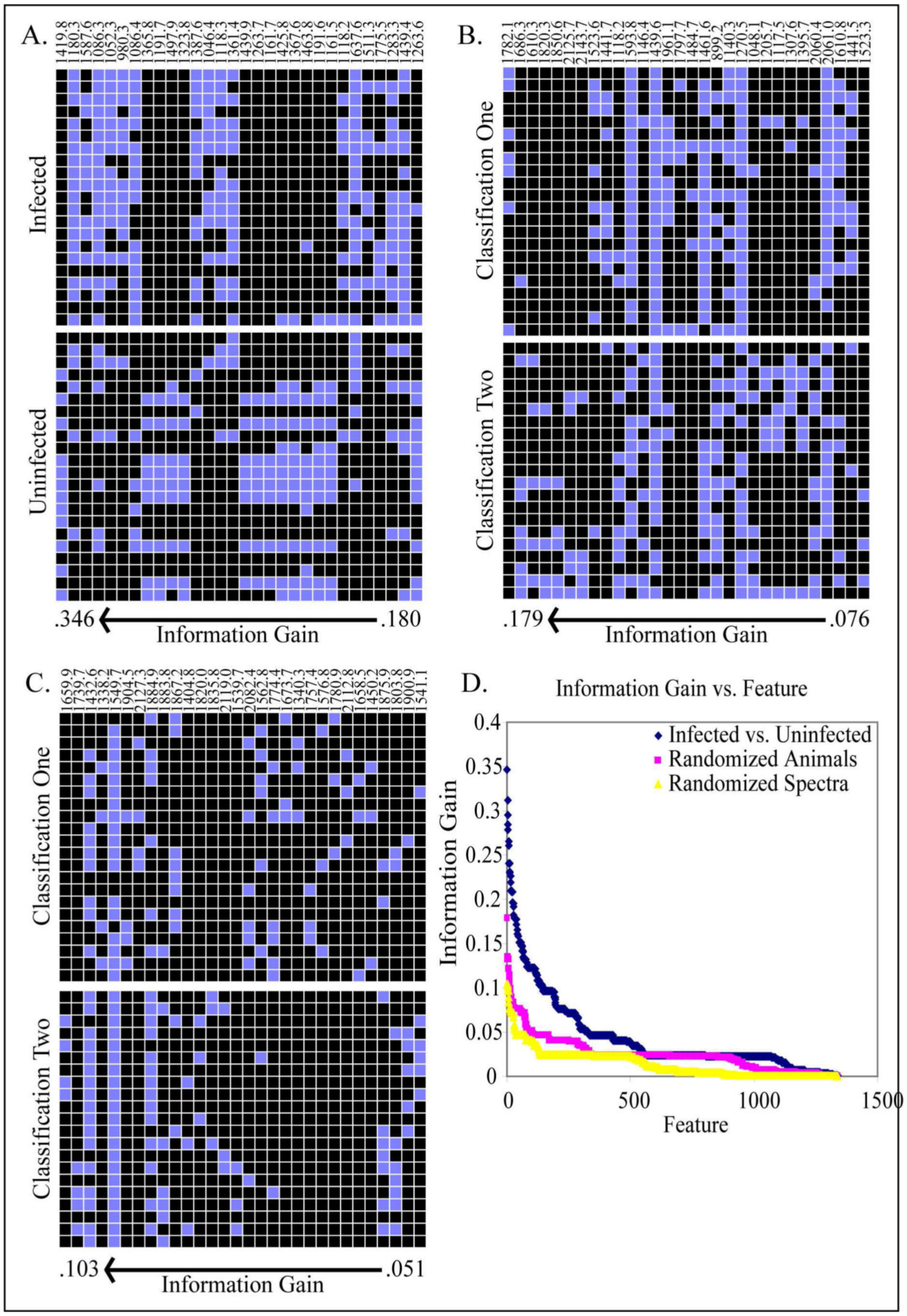


**Figure 2.**

Feature selection by information gain. A, B, C, heat maps displaying the top 30 bins ranked by information gain scores. A blue pixel indicates the presence of a peak in three out of the nine subsamples from each animal. A. Top 30 features distinguishing infected from uninfected animals. B, C, Samples are randomized at the animal (B) and spectra (C) level and the top 30 features are sorted by information gain. D. Plot of all 1,500 bins and their associated information gain score. Note that classification based upon disease state yields the best information gain scores.

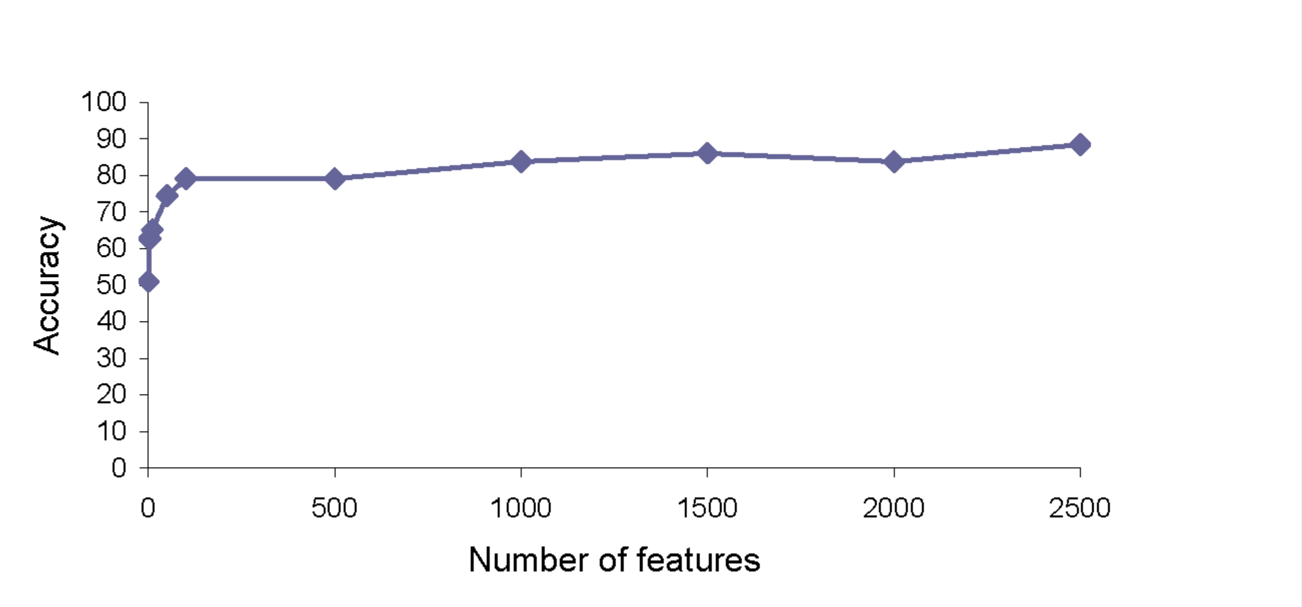
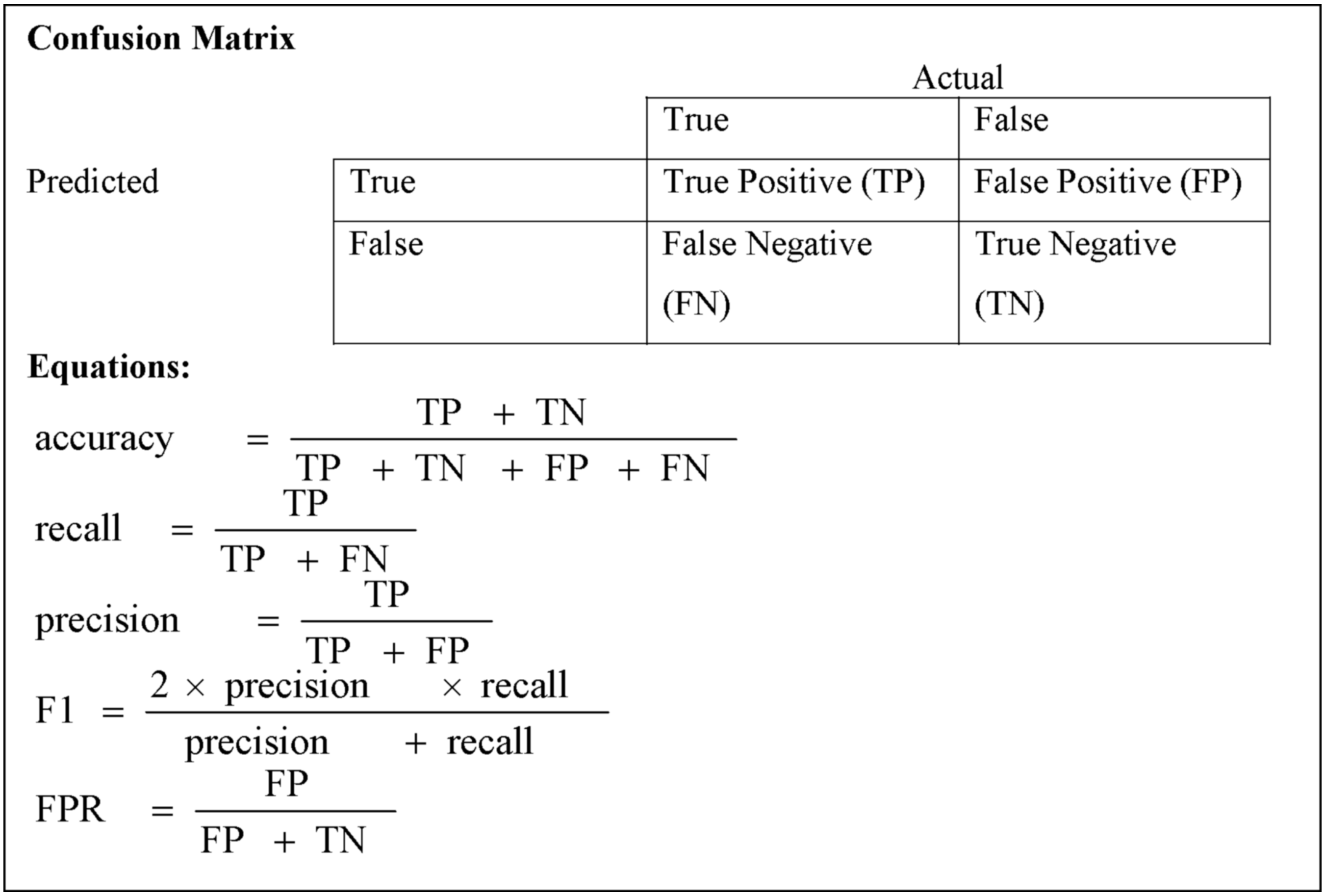


**Figure 3.**

Feature number versus accuracy. As the number of features incorporated into the model increases, the accuracy of the disease class predication improves up to ~100 features. After that, additional features provide no increase in overall accuracy.



**Confusion Matrix**

| | | Actual | |
|---|---|---|---|
| | | True | False |
| Predicted | True | True Positive (TP) | False Positive (FP) |
| | False | False Negative (FN) | True Negative (TN) |

**Equations:**

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$FPR = \frac{FP}{FP + TN}$$

**Figure 4.**

Confusion Matrix and performance equations. A confusion matrix has four measurements as follows: True Positive – Sample is actually positive and correctly predicted as positive. True Negative – Sample is actually negative and correctly predicted as negative. False Positive – Sample is actually negative and incorrectly predicted as positive. False Negative – Sample is actually positive and incorrectly predicted as negative. From this confusion matrix we can now describe a number of overall performance measures: Accuracy – Overall predictive accuracy of model, Recall – Ratio of positives correctly predicted, also sensitivity, Precision – Of positives predicted, how many were actually positive, F1-Score – 1st moment between precision and recall, False Positive Rate – Ratio of false positives predicted (specificity is 1-FPR).

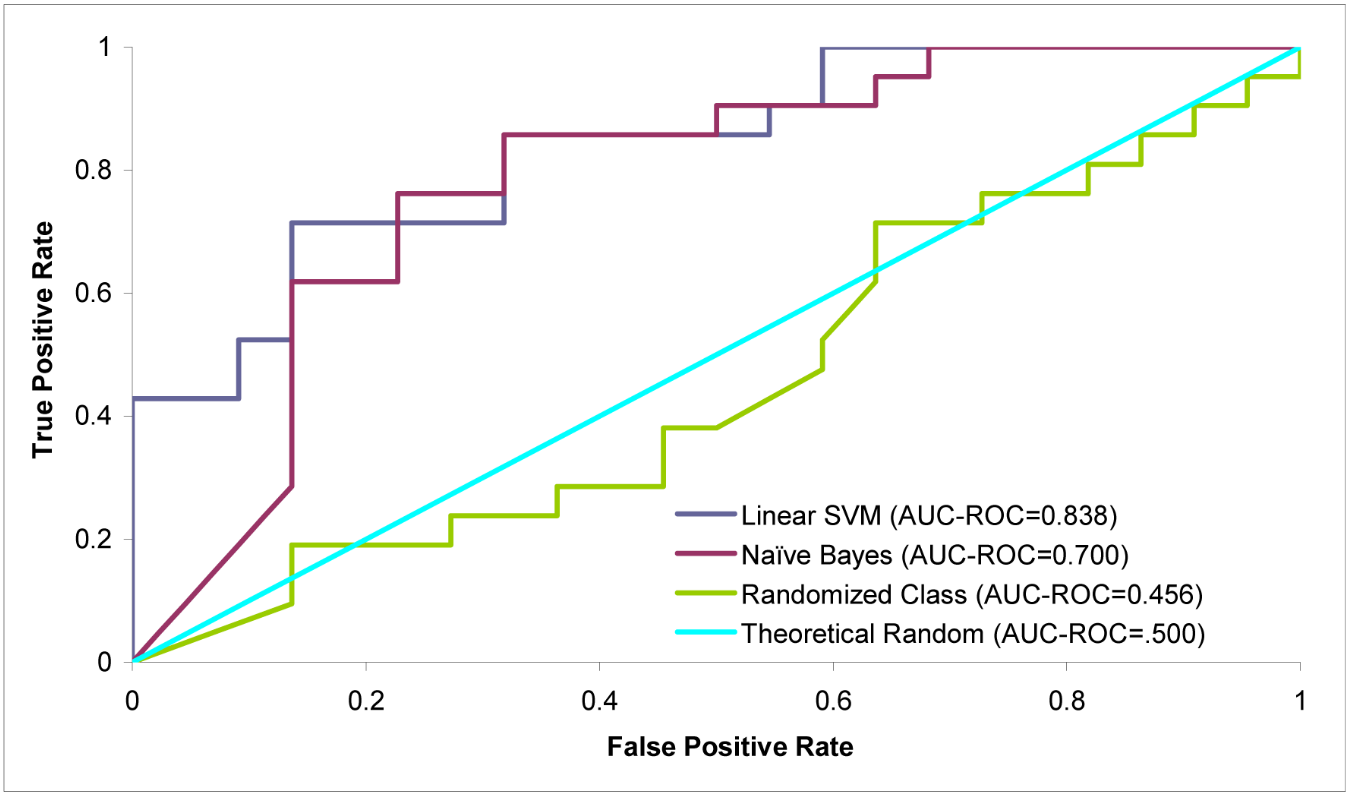


**Figure 5.**

Receiver Operator Characteristic Curve. Plotting the true positive rate (sensitivity) as a function of the false positive rate (specificity) allows visualization of the trade off between the two. The light blue line is the line of no discrimination, a derived from a theoretically random classification. The classification models for discriminating disease state (linear SVM and Naïve Bayes) reside in the upper left quadrant indicating a signal of disease. The light green line is derived from a randomization of the data set and “dances” around the theoretical random line indicating a high level of noise and an inability to discriminate between the two classes.

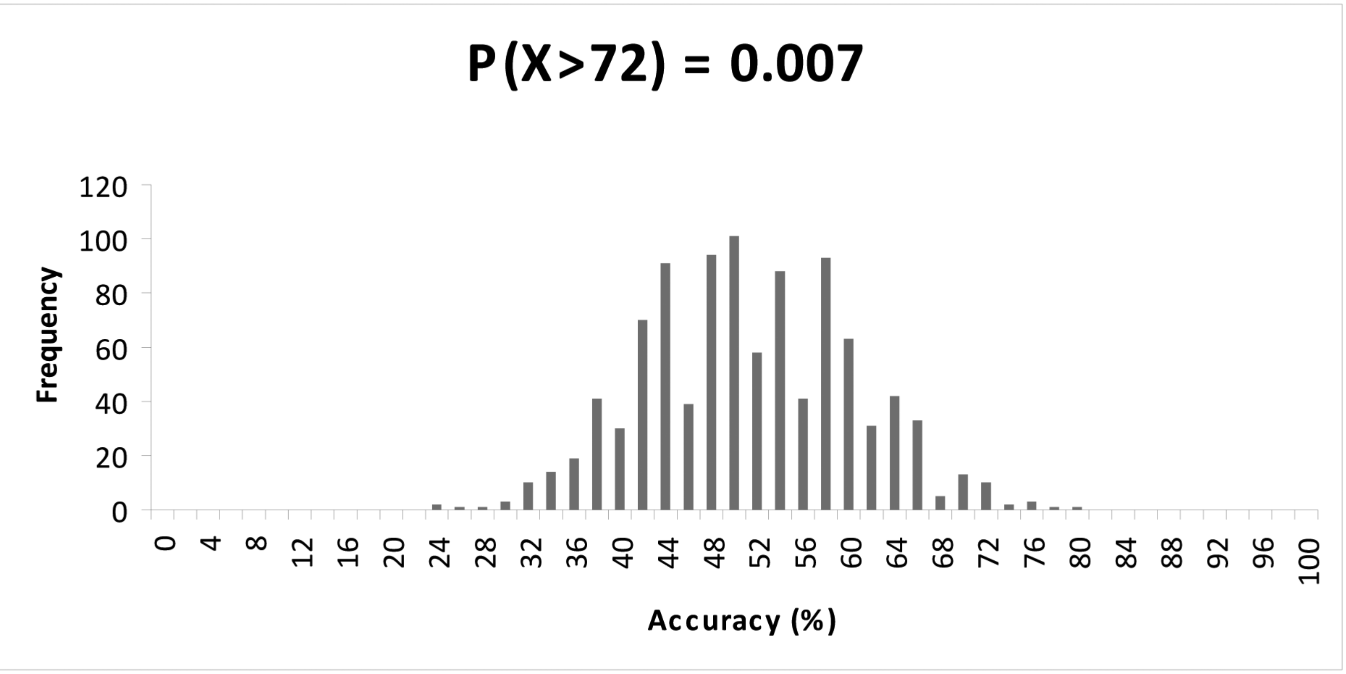


**Figure 6.**

Permutation test of SVM classifications. One thousand arbitrary classifications were performed on one thousand permuted data sets generated by randomization. The plot shows the predictive accuracy of each classification obtained by attempting machine learning on the randomized data sets and estimates the probability of obtaining a predictive accuracy of 72% (p<0.007), the accuracy at which prion disease was able to be distinguished from control.