Differentiating *Shigella* from *E. coli* using hierarchical feature selection on MALDI-ToF MS data

Author:
Laurens van de Wiel

Supervisor Radboud:
Prof. Dr. T.M. Heskes
Supervisors TNO:
Dr. A. Paauw
Dr. E. Tsivtsivadze

Radboud University Nijmegen

TNO

Master's Thesis
Computing Science
Radboud University Nijmegen
January, 2014

Abstract

Shigella is a genus of pathogenic enteric micro-organisms, which in some cases pose a lethal threat in a human host. It is closely related to *E. coli*, another pathogenic enteric micro-organism. Although in a clinical observation they cause different symptoms, on a genetic level they are very similar (Brenner, Fanning, Miklos & Steigerwalt, 1973). This poses a problem when the need arises to differentiate the *Shigella* and *E. coli* genetically. A possible way to accomplish this, is through a MALDI-ToF MS analysis.

Differentiating *Shigella* from *E. coli* is important since *Shigella* is one of the sources to infectious diarrhea that cause a problem in both developing and developed countries worldwide, that has a potential lethal result (Cheng, McDonald & Thielman, 2005).

Our work is based on MALDI-ToF MS data containing a comprehensive analysis of various *Shigella* species and *E. coli* phylotypes. We propose an approach which leads to a proper differentiation between both *Shigella*, *E. coli* and their respective species or phylogroups.

We make use of the elastic net method (Zou & Hastie, 2005) to extract the most important features that allow this differentiation without losing any correlation between the features.

We further extend the elastic net method to be build-up in using multiple models abiding a hierarchical structure in order to increase prediction performance. We compare two hierarchical structures, one based on the evolutionary phylogeny and one on pathotype (Y. Zhang & Lin, 2012). Using a hierarchy we can increase predictive performance and it allows us to find features specific for each hierarchy level, without having any added noise of unrelated hierarchy levels.

Furthermore we compare the impact in terms of predictive performance, feature selection and feature stability, by using two approaches to preprocessing the raw MALDI-ToF MS data. In both approaches we use common data-preprocessing techniques, but in one case we leave out the aggressive data reduction steps, smoothing and binning.

1 Introduction

Infectious diarrhea is a problem in both developing and developed countries worldwide, and accounts annually for more than 2 million deaths (Cheng et al., 2005). *Shigella* is an enteric bacterial pathogen, leading to possible lethal casualties, and is one of the sources of this problem.

A fast, accurate, inexpensive, and specific identification of enteric bacterial pathogens is considered essential for a proper therapy of diarrheal illnesses (Bennett Jr & Tarr, 2009). However, the conventional methods are time-consuming and labor-intensive, and they require experienced clinical microbiologists (He, Li, Lu, Stratton & Tang, 2010). Furthermore differentiating *Shigella* from *E. coli* proves a challenging task for many of modern day analysis methods (He et al., 2010).

MALDI-ToF MS[1] is a fast, accurate, and inexpensive method used for analysing biomolecules (e.g. DNA, proteins and, peptides). The corresponding data may be used for various applications.

The MALDI-ToF MS process uses a time measurement for an ion (i.e. signal wave) to travel along a flight tube to a detector. The duration of such an ion travelling within this flight tube is affected by the mass and charge of the ion. This time-representation can be translated into mass to charge ratio (m/z) over a range of Da[2] and is a translation of the mass of the analyte. The resulting data is represented as a *spectrum* using the relative intensity values at various m/z intervals (Tong et al., 2011). Such a spectrum is characterized by its peaks in intensity, which may contain signatures identical to a specific profile (Deshpande et al., 2011). We observe such peaks as the *features* that correspond to a specific biological sample, where each peak is denoted by their intensity value based on a specific m/z point.

The MALDI-ToF MS process detects the most abundant proteins, which are mostly ribosomal. Such abundance may be directly translated from the peak intensities and their locations present in the spectra. The peaks may therefore be considered as *biomarkers* in the form of proteins, which provide important information about biological questions. Such questions may entail the pathological state of certain diseases, disease progression, or species identification(X. Zhang, Zhu, Xiong, Deng & Zhang, 2013).

Biomarkers are especially important as they may be used for new biological insights about the relations within the complex data as well as reduce the data into fractions of what is important for a specific problem. Identifying and/or discriminating these biomarkers, is called *biomarker identification* (Deshpande et al., 2011). Throughout this paper we will refer to biomarker identification as *feature selection*, since there is no difference between both terms except for the fact that a biomarker denotes a feature in a biological context.

Raw MALDI-ToF MS data further contains a large degree of noise which is due to its sensitivity. For example, mechanical noise caused by the instrument settings or chemical noise that is influenced by sample preparation (Tong et

[1] Matrix-assisted laser desorption ionization time-of-flight mass spectrometry
[2] Da is the 'unified atomic mass unit' or Dalton

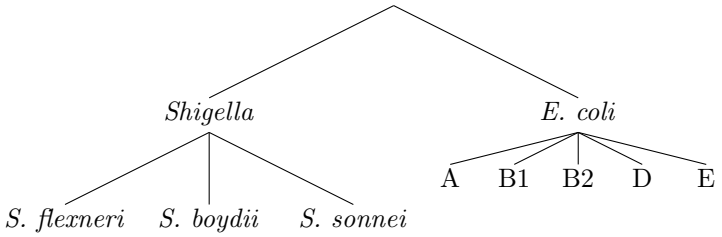

Figure 1.1: The hierarchical structure present in the data, based on the evolutionary phylogeny between *E. coli* and *Shigella* (Y. Zhang & Lin, 2012)

al., 2011). Due to this high susceptibility of noise, data preprocessing is often considered an important step to reduce bias and noise from the data before a proper analysis may be made (Cannataro, Guzzi, Mazza & Veltri, 2005).

Preprocessing of MS data ensures comparable spectra and may use data reduction techniques in order to reduce the dimensionality of the features. However, during such data reduction techniques important features may be lost (Coombes, Baggerly & Morris, 2007). For this reason we have made a comparison between two datasets resulting from an approach with a 'standard' preprocessing approach, and one which leaves out these data reduction techniques. The comparison entails on how data reduction impacts predictive performance (see Section 4) and the impact of data reduction on feature selection, as discussed in Appendix A.

We have made use of the Elastic Net method (Zou & Hastie, 2005) to allow both a feature selection to occur while performing a proper differentiation. We further extend the elastic net method to be build-up using multiple models abiding a hierarchical structure in order to increase prediction performance.

1.1 Problem setting

The data (Section 2) consists of MALDI-ToF MS analyses on 3 species of *Shigella* and 5 phylogroups of *E. coli*.

We are interested in finding the most important features within this data for the following sub-problems:

1. Features that allow discrimination between the *Shigella* and *E. coli* genera.
2. Features that allow the discrimination on a species level for the *Shigella* genus.
3. Features that allow the discrimination on a phylotypic level for the *E.coli* genus.

Each of these problems may be posed in a classification setting and solved by the elastic net. However, since we are dealing with a biological question, there exists a certain hierarchy between the classes. Specifically between *Shigella* and *E. coli* there exists a hierarchy based on evolutionary phylogeny (Fig 1.1) and one based on pathotype (Fig 1.2).

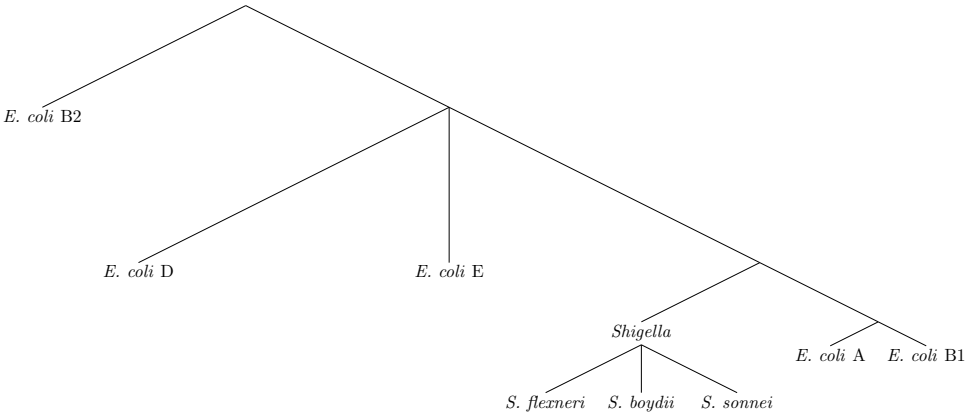

Figure 1.2: The hierarchical structure present in the data, based on the pathogenic similarity between *E. coli* and *Shigella* (Y. Zhang & Lin, 2012)

We can make use of this hierarchical knowledge by building multiple models of the elastic net, each specifically for a hierarchy node. This way all the subproblems may be solved in one single method. Such a model build-up generally has a better predictive performance than a flat multi-class classification if a hierarchy exists (Secker et al., 2007). Also creating elastic net models specifically per hierarchy node allows us to find features that are specific for that level, without having any added noise of unrelated hierarchy levels.

A full description on how we approach these problems and the methods involved is described in Section 3. The results obtained using a multi-class setting and the hierarchical model may be found in Section 4 followed by the conclusion in Section 5. In Appendix A we focus on how well suited data reduction steps in preprocessing are, in terms of feature extraction.

1.2 Related Work

The suitability of using the MALDI-ToF MS method for analysing enteric bacterial pathogens is shown in the work by He et al. (2010). However, in their experiments they failed to effectively identify both the *Shigella* species and EHEC (a specific set of *E. coli* phylotypes) isolates using the 'Biotyper system'. This emphasizes the need for a specific identification method. Furthermore their method does not provide the identification of the specific biomarkers on which the genera or species are identified.

In the work by Deshpande et al. (2011), a software suite is developed for the identification of pathogenic micro-organisms within MALDI-ToF MS data. Different to our problem, this suite focusses on identifying different genera instead of a specific species or phylogroup. They use an unsupervised approach in finding the relatedness between the test and database organisms and do not provide biomarkers.

Unsupervised hierarchical clustering has been used on a similar problem setting in the work by Hsieh et al. (2008). In this work they try to identify the hierarchy

between six bacteria genera including *E. coli* (but not *Shigella*) from MALDI-ToF MS data. Without providing any prior knowledge on how this hierarchy looks like, their method is able to exactly find the known hierarchy between these genera. This supports the fact that hierarchical relations are present within the MALDI-ToF MS data of *E. coli*.

In the recent work by Khot and Fisher (2013) a proper differentiation has been made between *E. coli* and *Shigella* based on MALDI-ToF MS data. However they have not incorporated any hierarchical knowledge in their model. Also the features they denoted as important where only so due to empirical testing, not the predictive performance that we obtain by using the elastic net. Furthermore they used data reduction techniques, which we show the negative impact of in Appendix A.

2 Dataset

Our dataset consists of raw and unprocessed 214 MALDI-ToF MS analyses of 135 *Shigella* and 79 *E. coli* samples. The *Shigella* samples further consist of 52 *S. flexneri*, 11 *S. boydii*, and 72 *S. sonnei* samples. The *E. coli* samples consist of the phylogroups[3], 25 A, 18 B1, 13 B2, 12 D and 11 E.

In these MALDI-ToF MS analyses, two different matrices[4] have been used. A matrix based on Ferulic Acid (FA+) and another matrix based on α-Cyano-4-hydroxycinnamic Acid (HCCA). These analyses each result in a spectrum that is identical for the matrix used. Furthermore, each analysis has been performed 4 times, resulting in 4 spectra per sample analysed.

To utilize the multiple analyses per sample, we average each of them for the corresponding sample. In this way we are able to filter out some of the noise specific to each analysis and put more focus on the common intensities. A single sample in our datasets is the averaged representation of 4 analyses of a single biological sample.

The HCCA matrix results in a spectrum accurate in the m/z range of 2000Da up to 12000Da. Where the FA+ matrix is accurate in the range of 8000Da up to 55000Da. This corresponds, without any preprocessing, to a potential of 10000 features for HCCA and 47000 features for the FA+ matrix. Since the HCCA and FA+ both have different spectra due to the different chemical agent used and the corresponding reaction to the biomolecules, we view them as two different datasets over the same samples.

[3]For an extensive analysis on the *E. coli* phylotype grouping we suggest reading the work by Jaureguy et al. (2008)

[4]A matrix in a MALDI-ToF MS setting entails a plate coated with a chemical agent that ionizes the biomolecules

3 Methods

3.1 Preprocessing

When dealing with raw MALDI-ToF MS data, preprocessing is required in order to form a proper dataset consisting of features comparable across all spectra. Preprocessing of MS data focusses on (1) noise reduction, (2) reduction of non-informative data and (3) allow different spectra to become comparable. This process may also be referred to as *peak detection*. The peak detection performed by Cannataro et al. (2005), Wagner, Naik and Pothen (2003), Coombes et al. (2007) and Yang, He and Yu (2009) all use the following order of approach:

1. **Base line identification**, attempts to identify the baseline between the many detected m/z intensity points.
2. **Peak alignment**, by selecting common reference peaks in the spectra, the different samples are aligned on these peaks by rescaling and shifting the spectra in order to fit to these reference peaks. This is needed due to the different intensities and their locations between the samples, which are influenced by the many different noise factors in MS analysis.
3. **Normalizing the intensities**, corrects any systematic differences between the samples.
4. **Smoothing of intensities**, attempts to smooth out any further noise between the samples.
5. **Peak extraction**, selects peaks between the average occurring peak width as the maximal value occurring above a certain minimal peak value. This may be seen as simply filtering out the peaks.
6. **Binning**, used to reduce the peaks from the peak extraction step by performing clustering. The cluster centroids of the found clusters are then further used as the peak locations.

There are a variety of available tools[5] that make use of these preprocessing methods on a given MS dataset. For the research conducted in this thesis, we made use of the the Matlab BioInformatics Toolbox (Schmidt & Jirstrand, 2006). Within this toolbox methods are available that allow construction of a tailor made solution respecting the above steps.

We have used the methods using their default settings in order to create two datasets per matrix, one that uses steps 1 to 6 and one that uses only steps 1, 2, 3 and 5. We refer to these datasets as the *binned* and the *raw* dataset (see Table 3.1). The combined dataset is constructed using the resulting datasets of separately preprocessing the HCCA and the FA+ matrix.

We have specifically chosen to not include the smoothing (step 4) and the binning process (step 6) for the raw dataset. We worked under the assumption that the elastic net method (Zou & Hastie, 2005) is able to cope with any redundant

[5] An overview of publicly available peak detection algorithms is provided in the work by Yang et al. (2009)

Matrix	# Features Raw	# Features Binned
FA+	18757	870
HCCA	5743	158
Combined	24500	1028

Table 3.1: Datasets constructed due to preprocessing.

and noisy features, therefore avoiding the possibility of losing important features that may get lost using the smoothing and binning techniques.

3.2 Feature selection

Feature selection focuses on finding the most important features. Depending on the question, the definition of what an important feature is may change. When focussing solely on features that are of an empirical significance, no question is posed on the meaning of these features. In such a case no predictions are needed and simple to more complex data analyses may be made. However, when interested in the important features under a certain context we may result to the feature's predictive performance. Doing so for a subset of features there are typically three flavours (Guyon & Elisseeff, 2003):

- **Wrappers**, utilize a machine learning method of choice in order to score subsets of features according to their predictive power.
- **Filters**, filtering out subsets of features as a pre-processing step independent of the chosen machine learning method.
- **Embedded methods**, perform feature selection in the process of training and often are specific to a machine learning method.

The elastic net method falls into the embedded method category.

3.3 The Elastic Net method

We consider the following situation where we have data $(\mathbf{x}^i, y_i)$, where $i = 1, 2, ..., N$ denoting the examples, and $\mathbf{x}^i = (x_{i1}, x_{i2}, ..., x_{ip})^{\mathbf{T}}$ as features for response y_i.

We assume that the examples are independent from each other and that the y_i's are conditionally independent given their corresponding $\mathbf{x}^i$.

We use the classical regression set-up with coefficients $\hat{\beta}$:

$$y_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \tag{1}$$

We also ensure that that each x_{ij} is standardized such that

$$\frac{\sum_i x_{ij}}{N} = 0, \qquad \frac{\sum_i x_{ij}^2}{N} = 1 \tag{2}$$

The elastic net method is a hybrid combination of the LASSO method (Tibshirani, 1996) and ridge regression (Hoerl & Kennard, 1970). A combination of these methods is used to provide the advantages that each method offers.

Ridge regression allows coefficients to shrink, providing more stability. However, the ridge regression is unable to shrink coefficients to zero. The LASSO method is able to shrink coefficients to zero by imposing an L_1-penalty on the regression coefficients, doing so allows an easily interpretable model by viewing any non-zero coefficients as the predictors that have the strongest predictive power to a posed problem. Also, overall prediction accuracy is improved with the LASSO method compared to the ridge regression by disregarding the bias from the zero-coefficients (Tibshirani, 1996).

A comparison on the predictive performance of the LASSO, ridge and other methods was made by Tibshirani (1996) and Fu (1998). They did not find that any of these methods was empirically better than the other. However, when considering feature selection as an important factor, due to the possibility of shrinking coefficients to zero, the LASSO is a much more proper choice.

Elastic net allows the benefit that is provided by the LASSO method, where coefficients for features may be shrunk to zero, hence provide interpretable models consisting of the most important features. The elastic net further has a similar performance as standard ridge regression (Zou & Hastie, 2005).

The elastic net overcomes the following limitations of the LASSO method (Zou & Hastie, 2005):

- In a setting where we have more predictors than samples (called the '$p \gg N$' problem) LASSO is able to have at most N non-zero coefficients. In the case of our MALDI-ToF MS data, we have '$N = 214$'. Thus the only dataset that would be suitable for the LASSO method is the binned HCCA, since that has '$p = 158$'. All the other datasets have more predictors than samples and as such would not be suitable for a LASSO approach (see Table 3.1).
- LASSO is not able to utilize correlation between coefficients, the LASSO tends to randomly select only one feature from a correlated group. By incorporating ridge regression as well, the elastic net allows the ridge regression tendency of shrinking coefficients for correlated features towards each other, while retaining the feature selection provided by the LASSO. This encourages a 'grouping effect', providing subsets of correlated features (Zou & Hastie, 2005). In the case of MALDI-ToF MS data, correlated features may be directly translated to correlated proteins. From a biological perspective we would not like to lose this information, since proteins and their correlations to other proteins may prove a valuable insight in their respective functions.
- In a situation where we have '$N > p$' and there exists high correlations

between coefficients, predictive performance of ridge regression tends to outperform the LASSO (Tibshirani, 1996).

The elastic net is often used for feature selection within MS data (F. Z. Zhang & Hong, 2011; Saeys, Inza & Larrañaga, 2007). The main motivation for choosing the elastic net method is to overcome the '$p \gg N$' problem, which is typical for MS data. Also the retaining correlation between features is often emphasized. Situations may occur where a single feature is useless by itself, but when combined with other features provide better predictive performance (Guyon & Elisseeff, 2003).

The elastic net model is described by the coefficients $\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p\right)^{\mathbf{T}}$. Zou and Hastie (2005) describe that we may leave out the *bias* term $\hat{\beta}_0$ from Equation 1. This is due to the definition of elastic net that when standardization is applied (see Eq. 2) the bias term may be set to zero and thus becomes obsolete.

The elastic net tries to minimize the the following problem in order to obtain estimates $\hat{\beta}$:

$$L(\lambda_1, \lambda_2, \beta) = |\mathbf{y} - \mathbf{X}\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|_1 \tag{3}$$

with

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2,$$
$$|\beta|_1 = \sum_{j=1}^{p} |\beta_j|.$$

When defining $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$ than the elastic net optimization problem is defined as:

$$\hat{\beta} = \arg\min_{\beta} |\mathbf{y} - \mathbf{X}\beta|^2, \quad \text{subject to } (1-\alpha)|\beta|^2 + \alpha|\beta|_1 \leq t \text{ for some } t. \tag{4}$$

The α-value allows control over the impact of L_1 and L_2 penalization. Elastic net can also be set to ridge regression ($\alpha = 0$) or LASSO ($\alpha = 1$).

For our elastic net implementation we have used the Scikit-Learn package (Pedregosa et al., 2011). The implementation of elastic net within Scikit-Learn writes the α variable from Equation 4 to consist of two elements. The Scikit-Learn optimization is subject to '$\mu(1-\rho)|\beta|^2 + \mu\rho|\beta|_1 \leq t$ for some t'. If we set $\mu = 1$ we obtain the Equation 4. The ρ and μ values are candidates for cross-validation further discussed in Section 3.6.

3.4 One vs all classification

We may use the elastic net method in a classification setting if we regard the classification problem as a binary regression problem. We do this by observing one class as '$y = 0$' and the other class as '$y = 1$'. We optimize during the cross-validation on the Area under the ROC curve (AUC) (see Section 3.6). Therefore, we cannot simply use a cut-off threshold value of '0.5' as you would when optimizing was performed on the accuracy for predicting the class to which an example belongs. When predicting the class we consider:

$$\hat{y_i} = \begin{cases} 1, & \text{if } \mathbf{x}^i\beta > \sigma. \\ 0, & \text{otherwise.} \end{cases}$$

Here σ represents the threshold value that is set in the following manner. Consider $\mathbf{X}'$ as the training set and $|\mathbf{y}'^{-}|$ as the number of examples that correspond to the label zero in the training set, then:

1. Cross-validate over the training set in order to retrieve the optimal parameters (see Section 3.6).
2. Create β' by fitting on the training set $\mathbf{X}'$ with responses $\mathbf{y}'$, using the optimal parameters from step 1.
3. Construct $\tilde{\mathbf{y}}'$ as the responses of $\mathbf{X}'\beta'$, predicting the responses for the samples from the training set.
4. Order the $\hat{\mathbf{y}}'$ from the lowest value to the highest value.
5. σ is equal to the $|\mathbf{y}'^{-}|$-ith occurring value in the ordered $\hat{\mathbf{y}}'$.

Although this may make σ prone to any over-fitting occurring on the training set, it is a more intelligent way to pick a threshold than plainly setting it to 0.5, and it respects the optimization occurring on AUC instead of accuracy.

When having multiple classes the above is further extended by transforming the classes using the following definition:

$$y_j^i = \begin{cases} 1, & \text{if } y_j = i. \\ 0, & \text{otherwise.} \end{cases}$$

This obtains the encoding matrix M where m_{ij} is the y value encoding for class j when constructing the model for class i (i.e. $M_{ij} = 1$ if $i = j$, otherwise $M_{ij} = 0$). For our case M is considered equal to the identity matrix I with a dimensionality of 7, corresponding to the 3 *Shigella* species and 4 *E. coli* phylogroups.

Such a setting for multiple classes requires the need to build a binary classification model '$\beta^{(i)}$' for each class $i \in \{1, 2, ..., C\}$ of C classes.

When M is set as the identity matrix, the multi-class prediction becomes:

$$\hat{y}_i = \arg \max_{c \in \{1,\ldots,C\}} \mathbf{x}^i \beta^{(c)} \tag{5}$$

This is called a One-vs-All (OVA) scheme (Rifkin & Klautau, 2004). In this setting there are trained as many binary classifiers as there are classes, and each classifier is trained according to their respective row in the M encoding.

3.5 Hierarchical Classification

The elastic net method has been used in the past for classification problems on MALDI-ToF MS data (Saeys et al., 2007). However, to the best of our knowledge, such approaches did not incorporate use of a hierarchical structure present within that data.

An OVA scheme proves to be as accurate as any other multi-class classification scheme, based on the assumption that the classes are independent of each other and do not belong to a natural hierarchy. When this is the case, an algorithm that exploits such relationships between classes could offer superior performance (Rifkin & Klautau, 2004).

Within our data we have two naturally occurring hierarchies. Between *Shigella* and *E. coli* there exists a hierarchy based on evolutionary phylogeny (Fig 1.1) and one on the pathotypes (Fig 1.2). In the field of hierarchical classification there are multiple ways to exploit these relationships present within the data. However, we require a feature selection to occur as well. Therefore, we require a hierarchical model that both abides the hierarchies present in the data and is able to provide a proper feature selection.

To this end we make use of the approach as presented by Secker et al. (2007). This approach uses the hierarchical knowledge to build multiple classification models, each specifically for a node in the hierarchy. Such a model build-up generally has a better predictive performance than a flat multi-class classification like the OVA scheme (Secker et al., 2007). Creating elastic net models per hierarchy node allows us to find features that are specific for that level, without having any added noise of unrelated hierarchy levels.

Representing the hierarchy in such a hierarchical model build-up from Figure 1.1 we refer to Figure 3.1 and for Figure 1.2 we refer to Figure 3.2. Such a predictive model is trained by only considering the examples of a training set that are relevant for the current hierarchy level. For example, all training examples are considered in the top of Figure 3.1. At the node 'OVA scheme over *Shigella* species' only the examples from the training set that are *Shigella* will be considered.

When predicting to what class an unknown example belongs, the topmost hierarchical classifier is used to denote the specific branch to which that example most likely belongs. This is followed by the same approach for each of the predicted branch's child nodes until a leaf is reached, that leaf represents the predicted class.

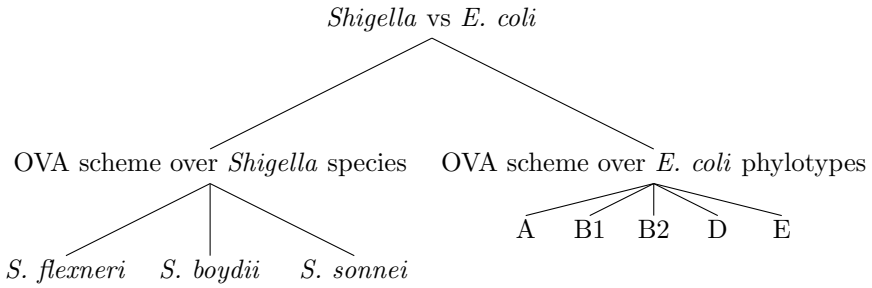

Figure 3.1: The hierarchical structure of Fig 1.1 set in a hierarchical classification scheme with the nodes as elastic net models and the leafs as classes

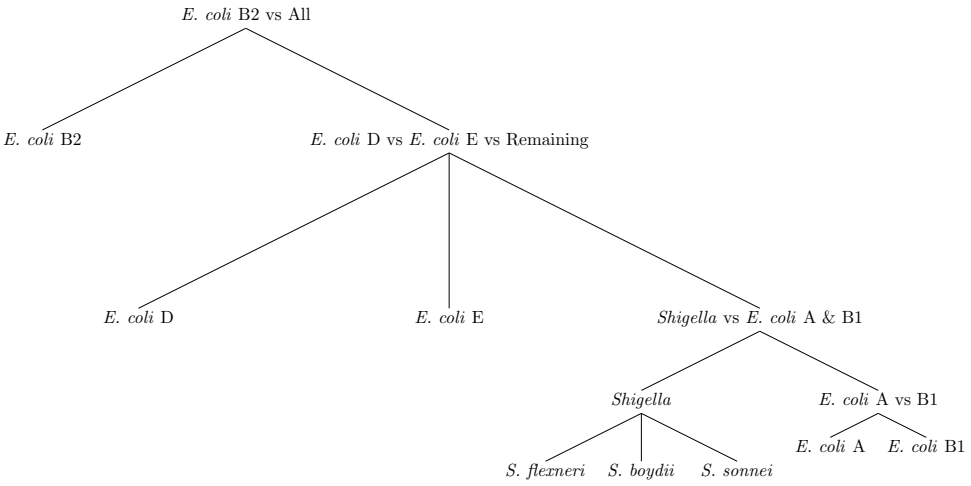

Figure 3.2: The hierarchical structure of Fig 1.2 set in a hierarchical classification scheme with the nodes as elastic net models and the leafs as classes

Using such a hierarchical classification scheme, we may obtain feature selection for each problem node, specific to only those of the related classes. In such a way we are able to gather more 'pure' features. Also this kind of scheme allows the elastic net to be used without any further adjustments.

3.6 Optimizing parameters

The results obtained in Section 4 have been achieved through a classical machine learning setting where we cross-validate on 70% of the data, called the *training set*, to find the optimal parameters for ρ and μ. We evaluate these optimal parameters found during cross-validation on the remaining and unseen 30% of the data, called the *test set*. Thus we achieve the final results.

To find optimal parameter $\alpha^* = (\rho^*, \mu^*)$ we perform a 5-fold cross-validation to find these parameters over the ranges:

$$\rho^* \in \{0.01, 0.1, 0.5, 0.7, 0.9, 0.95, 0.99\}$$
$$\mu^* \in \{e^{-4}, e^{-3}, e^{-2}, e^{-1}, e^{0}, e^{1}, e^{2}, e^{3}, e^{4}\}$$

In cross-validation we make use of a simulated train-test setting, where we split the training set into 5 equal folds. Then a training set is constructed using 4 out of the 5 folds where we try to retrieve the minimized set of coefficients β (see Eq. 4) using a single combination of ρ and μ values. β is then evaluated on the remaining fold where we make use of a metric that determines performance. This is repeated until every fold has been used as a test fold, achieving an average performance over the 5 folds.

This average performance is used to find the most optimal combination of parameters that fit the training set. Therefore, this process is further repeated until every parameter combination has been evaluated. This provides an indication of the most optimal parameters.

Different from the approach by F. Z. Zhang and Hong (2011), we optimize the elastic net in the cross-validation on the AUC. Since we are making use of the OVA scheme for multi-class classification, each classification problem may be considered as one or more binary classification problems.

For classification problems, optimizing on AUC is a better approach than optimizing on classification accuracy (Huang & Ling, 2005). Furthermore, optimizing on AUC is empirically equivalent to optimizing on the error rate (Cortes & Mohri, 2004). Although this may be influenced by the type of problem.

Dataset	Mean AUC (CV)	Mean AUC (Test)
Combined dataset	0.8831	0.8977
FA+ dataset	0.8394	0.7827
HCCA dataset	0.9339	**0.9219**

Table 4.1: AUC values from cross-validation and Test for OVA strategy for the binned datasets

Dataset	Mean AUC (CV)	Mean AUC (Test)
Combined dataset	0.7957	**0.8258**
FA+ dataset	0.734	0.7971
HCCA dataset	0.837	0.8058

Table 4.2: AUC values from cross-validation and Test for OVA strategy for the raw datasets

4 Results

Although AUC is a proper metric for optimization (Huang & Ling, 2005), we are unable to properly evaluate the AUC performance occurring in the hierarchical classification.

The AUC metric is not applicable to the hierarchical model since we may be evaluating samples that are not part of the hierarchy branch. Such an event may occur when a miss-classification happens earlier in the tree. For example an *E. coli* sample has been classified as *Shigella* in the model depicted in Figure 3.1. If this occurs than the *E. coli* will be classified as one of the *Shigella* classes, hereby polluting the AUC metric since it no longer ranges over classes that were actually in the scope when it was trained.

We do have an indication of the performance from the optimal parameters resulting from the cross-validation evaluated on the test set, but it only entails the OVA schemes. For the binned datasets we refer to Table 4.1 and for the raw datasets we refer to Table 4.2. These AUC values have been computed by taking the mean AUC from the AUC achieved for each of the 'one-vs-rest' folds. We may see that for the binned datasets the HCCA achieves the best AUC and for the raw dataset the combination of both the HCCA and FA+ is most optimal.

In order to compare if any improvement occurs due to using a hierarchical model build-up. We evaluate the accuracy and F_1-score of a hierarchical classification which is only based on the final class prediction, in this way we are able to compare the OVA scheme to the hierarchical classification. These results may be found in Table 4.3 for the binned datasets and in Table 4.4 for the raw datasets. We see that applying a hierarchical model build-up based on the evolutionary phylogeny (Fig. 3.1) proves to be the most optimal in most cases. Only on the binned HCCA dataset the OVA is a clear winner, but when comparing it to the

Dataset	# Features	Strategy	Accuracy	F_1-score
Combined dataset	1028			
		OVA	0.708	0.72
		H-Phylogeny	**0.723**	**0.73**
		H-Pathotype	0.415	0.46
FA+ dataset	870			
		OVA	0.569	0.59
		H-Phylogeny	**0.646**	**0.66**
		H-Pathotype	0.415	0.46
HCCA dataset	158			
		OVA	**0.754**	**0.74**
		H-Phylogeny	0.692	0.68
		H-Pathotype	0.662	0.69

Table 4.3: Results from binned datasets

binned combined dataset they are not very different.

We may also see that the hierarchy based on the pathotype (Fig. 3.2) does not perform well. This may have the cause that the hierarchy based on pathotype is not highly represented in the MALDI-ToF MS data, therefore the differentiating problems become increasingly more difficult when descending into the hierarchy. From our own observation we have observed that in the case of the raw Combined and the raw FA+ datasets all test examples are classified as *E. coli* B2, showing that in the top hierarchy node many miss-classifications occur.

Dataset	# Features	Strategy	Accuracy	F_1-score
Combined dataset	24500			
		OVA	0.462	0.51
		H-Phylogeny	**0.538**	**0.57**
		H-Pathotype	0.062	0.01
FA+ dataset	18757			
		OVA	0.215	0.20
		H-Phylogeny	**0.462**	**0.49**
		H-Pathotype	0.062	0.01
HCCA dataset	5743			
		OVA	**0.523**	**0.58**
		H-Phylogeny	**0.523**	0.53
		H-Pathotype	0.246	0.30

Table 4.4: Results from Raw datasets

5 Conclusion

Results show that utilizing an existing hierarchy generally improves the predictive performance compared to the OVA scheme. Even more, we have been able to properly differentiate *Shigella* and their species from *E. coli* and their phylogroups. These results make the MALDI-ToF MS analysis a proper method to use for differentiating between both genera.

In Appendix A we show that when the interest is in finding the most important features within MALDI-ToF MS data, excluding the data reduction steps during preprocessing is a more proper approach.

The features from a binned dataset cannot be directly mapped back to a specific peak in the spectrum. This is due to the clustering of the smoothed peaks that occurs during the binning process. Because of this an important feature in the binned datasets is nothing more than an indication of one or more biomarkers occurring near a specific m/z point. This is not the case when dealing with the raw datasets, where we may directly map the features back to the peaks.

In future studies that use a MALDI-ToF MS analysis in order to differentiate *Shigella* from *E. coli*, we would recommend using the combination of both the HCCA and FA+ matrix. In current studies, often only the HCCA matrix is used and is considered to be the 'standard' (He et al., 2010). However, our results show that a combination of both matrices has a performance that is close to solely using HCCA. Although FA+ by itself generally has a lower performance, in the combined dataset features are found from both matrices (see Appendix A). Indicating important features occurring in both matrices. Furthermore a combination of the matrices allow peaks to correlate across both matrices.

5.1 Future work

Results show that applying a hierarchical model construction improves the predictive performance. This improvement may be further increased using a global hierarchical classification technique. Such a technique like the work of Qiu, Gao and Huang (2009), requires a penalty to be put as a regularization term that corrects the miss-classifications within hierarchies. To the best of our knowledge this is a novel approach when using the elastic net method.

The reason we have not yet pursued this idea is due to the fact that the feature selection was the most important result needed from this research. When incorporating the hierarchy in the regularization term it puts a bias on the resulting coefficients, by being influenced by the hierarchy. Also, the current approach provides feature selection to occur per each hierarchy node.

Using the suggested global approach does benefit in the sense that it is less prone to miss-classification due to wrong classifications in higher hierarchy.

Acknowledgements I would like to thank the TNO research institute for allowing me to conduct all the work performed on this thesis and providing me with the data used for the analyses. A small thank you is in place for Wessel van Staal for helping me form the final look of the title page of my thesis. A very special thank you goes to my supervisors Armand Paauw, Tom Heskes, and Evgeni Tsivtsivadze. All of them proved to be a very valuable discussion partner, they allowed me to use their time in order to form the proper direction to go with this thesis and above all I really enjoyed collaborating with them.

Preprocessing	AUC (CV)	AUC (Test)
Binned	0.9770	**0.9772**
Raw	0.9192	0.9643

Table A.1: *Shigella* Vs *E. coli* AUC values for both the raw and binned combined dataset

A Case study: Comparison of feature selection in Binned and Unbinned MALDI-ToF MS data

In this case study we look into differentiating *Shigella* from *E. coli*. For this differentiation we are interested in finding the most predictive features (i.e. the biomarkers) and for this purpose we use stability selection (Meinshausen & Bühlmann, 2010). We evaluate the difference between the two approaches to preprocessing (see Section 3.1). Where the binned dataset is constructed using smoothing and binning in order to reduce the amount of data. The raw dataset is constructed using the same approach as the binned dataset, however, the data reduction steps are excluded. During this case we focus solely on the dataset that is combined from the HCCA and FA+ matrix (see Table 3.1).

In order to obtain the most predictive features, we first perform cross-validation to find the optimal pair value for α^* for each dataset (see Section 3.6). The results from the cross-validation performance and the results of evaluating α^* on the test set may be found in Table A.1. We observe that the binned dataset performs slightly better on the test set than the raw dataset in terms of AUC.

To find the most predictive features we use α^* in order to perform stability selection. Stability selection filters out features that are the most stable when constructing a model given a set of regularization parameters, in our case α^*. Stability denotes how likely a feature is to be used in the construction of a model. The more stable a feature is, the more likely is has a predictive performance on the posed problem (Meinshausen & Bühlmann, 2010).

The stability of a feature i is defined as follows (Meinshausen & Bühlmann, 2010):

$$\Pi_i^{\alpha^*} = P^*(i \in S^{\alpha^*}(V)) = \frac{1}{|V|} \sum_{v \in V} 1_{\{i \in S^{\alpha^*}(v)\}} \tag{6}$$

Here V is a set of random sub-samples from the dataset, each $v \in V$ is of size $\lfloor \frac{N}{2} \rfloor$ and S^{α^*} is called the structure estimate for the set of samples 'v', consisting of the non-zero coefficients when fitting on the set of samples (i.e. the relevant features for a single model).

A feature i is to be considered as stable, when $\Pi_i^{\alpha^*} \geq \epsilon$. Here ϵ is a cut-off threshold in the range $[0, 1)$. We specify '$\epsilon = (\max_{i \in \{1,2,\ldots,p\}} \Pi_i^{\alpha^*}) - 0.1$'. This

means that the feature that has the highest stability is selected, and all features that have at most a 10% difference to this feature are selected as well to form the set of most stable features. For $|V|$ we use the suggested value of $|V| = 100$ (Meinshausen & Bühlmann, 2010).

The stability selection is performed using 100% of the data (recombining the training and test set). So subsets ($v \in V$) may be formed from the full dataset. The optimal parameter α^* is still attained by a 70% and 30% split on the dataset.

In the binned dataset we find 307 features occurring at least once in the model construction (Fig. A.1) out of a total of 1080 possible features (Table 3.1), this accounts for 28.4% of possible biomarkers and a noise reduction of 71.6% to the total feature range for the posed problem.

In the raw dataset we find 2004 features occurring at least once in the model construction (Fig. A.2) out of a total of 24500 possible features, this accounts for 8.18% of possible biomarkers and a noise reduction of 91.82%. From this we may conclude that the elastic net is able to significantly reduce noise from a large set of features.

When observing the shape of the curve, we see that the binned dataset is more formed as a steep slope than the raw dataset, which has a concave slope. Typically we would rather see a concave slope, since this makes the 'peaked' features that are eventually denoted as stable features to have a greater importance.

The steep slope may be caused due to the binning process. The binning process combines multiple peaks into a single cluster. This way important features could become joined together. The joining may cause an increase in predictive performance, as we may see in the results section. However, when interested in the set of best performing features in the context of MALDI-ToF MS data, we would like to exactly know the location of the peaks.

One thing that may occur in the raw dataset, is a feature to be selected that is very close to another feature. This may implicate two things, either we are dealing with two separate protein peaks, or we see a shoulder peak occurring that is correlated with the other feature. Due to the elastic net's grouping effect, we will extract both of these features. If indeed an important feature is a shoulder peak, we need to check this with the actual spectrum.

Also the most stable features found are far more abundant in the case of the binned dataset, where we have 44 stable features (Fig. A.3). For the raw dataset we have 18 stable features (Fig. A.4). Considering how many features occur at least once in the stability selection process, the raw dataset selects 18 out of 2004 features, compared to 44 out of 307 for the binned dataset. This seems to make the found stable features of a far greater importance in the case of the raw dataset, especially given that the original scope was 24500 features.

Comparing the performance between the optimal parameters found for α^* on the binned and on the raw dataset, there is only a slight difference in AUC when evaluating on the test set. Also, given the shape of the stability curve, the amount of features denoted as stable, and the fact that binning makes it harder to retrace the exact position of the important peaks in the MALDI-ToF MS

spectrum, we would defend that binning should not be used when interested in finding the most important features for a problem.

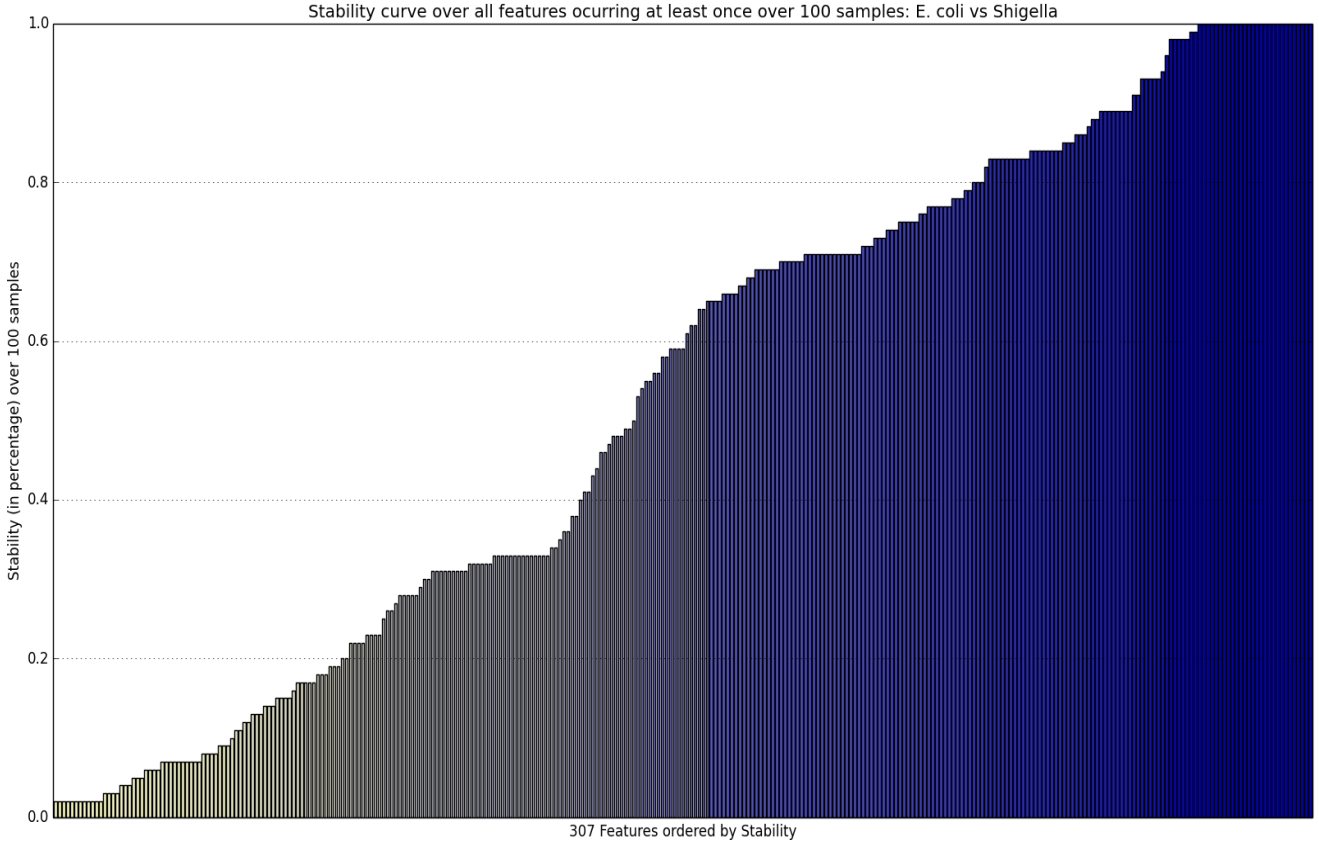

Figure A.1: The stability curve of features from the binned dataset occurring at least once in the model construction over 100 sub-samples.

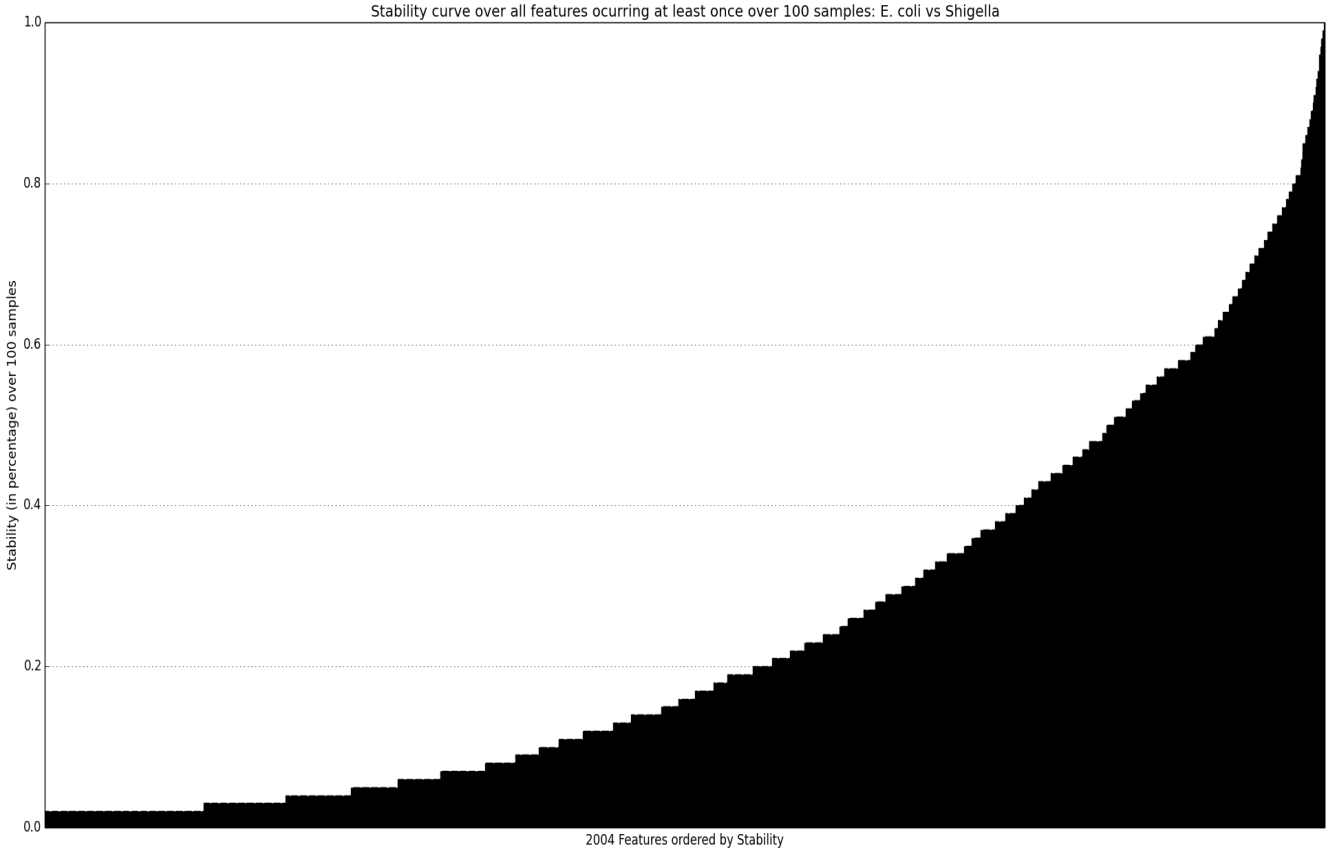

Figure A.2: The stability curve of features from the raw dataset occurring at least once in the model construction over 100 sub-samples.

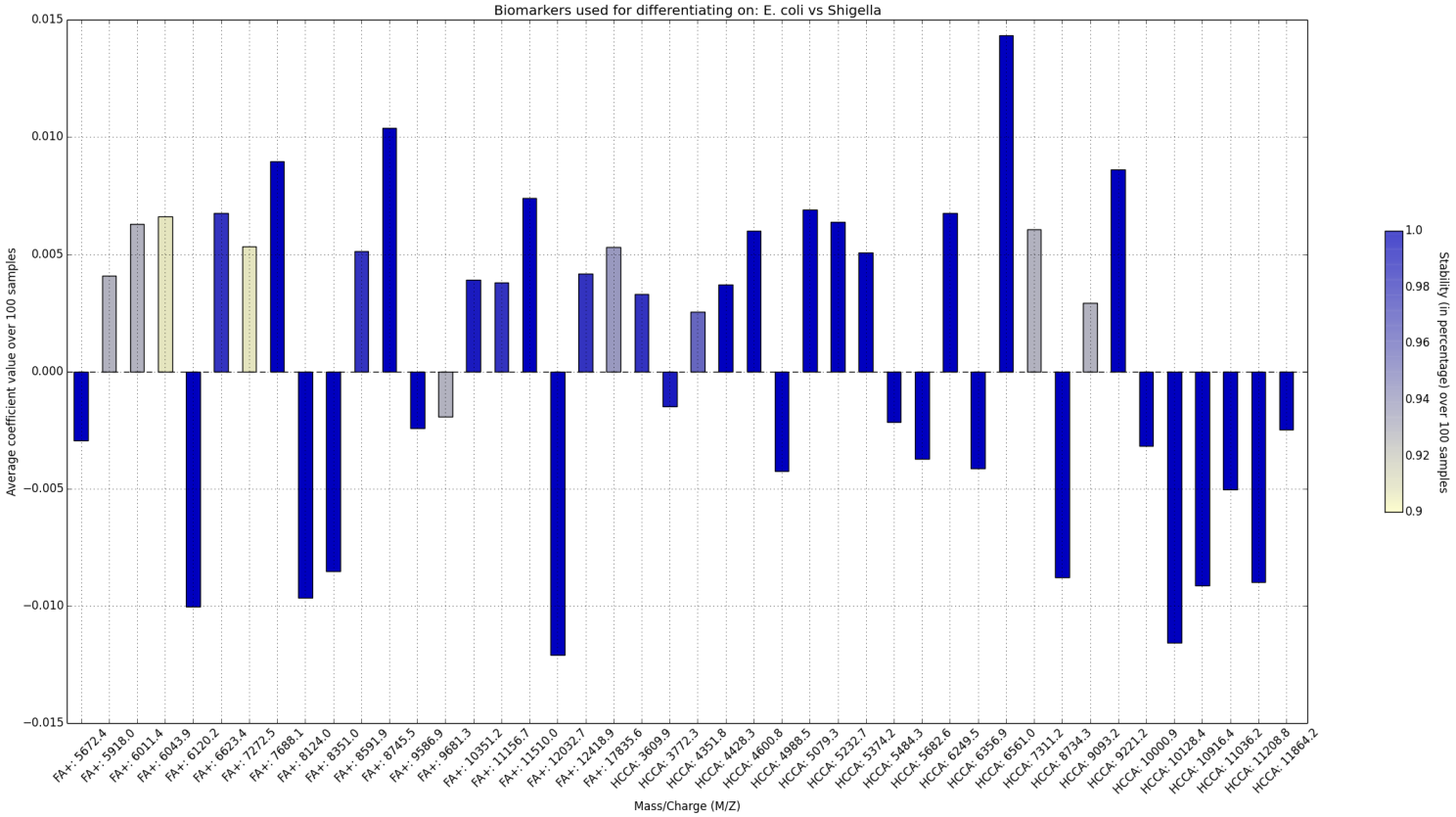

Figure A.3: The most stable features from the binned dataset.

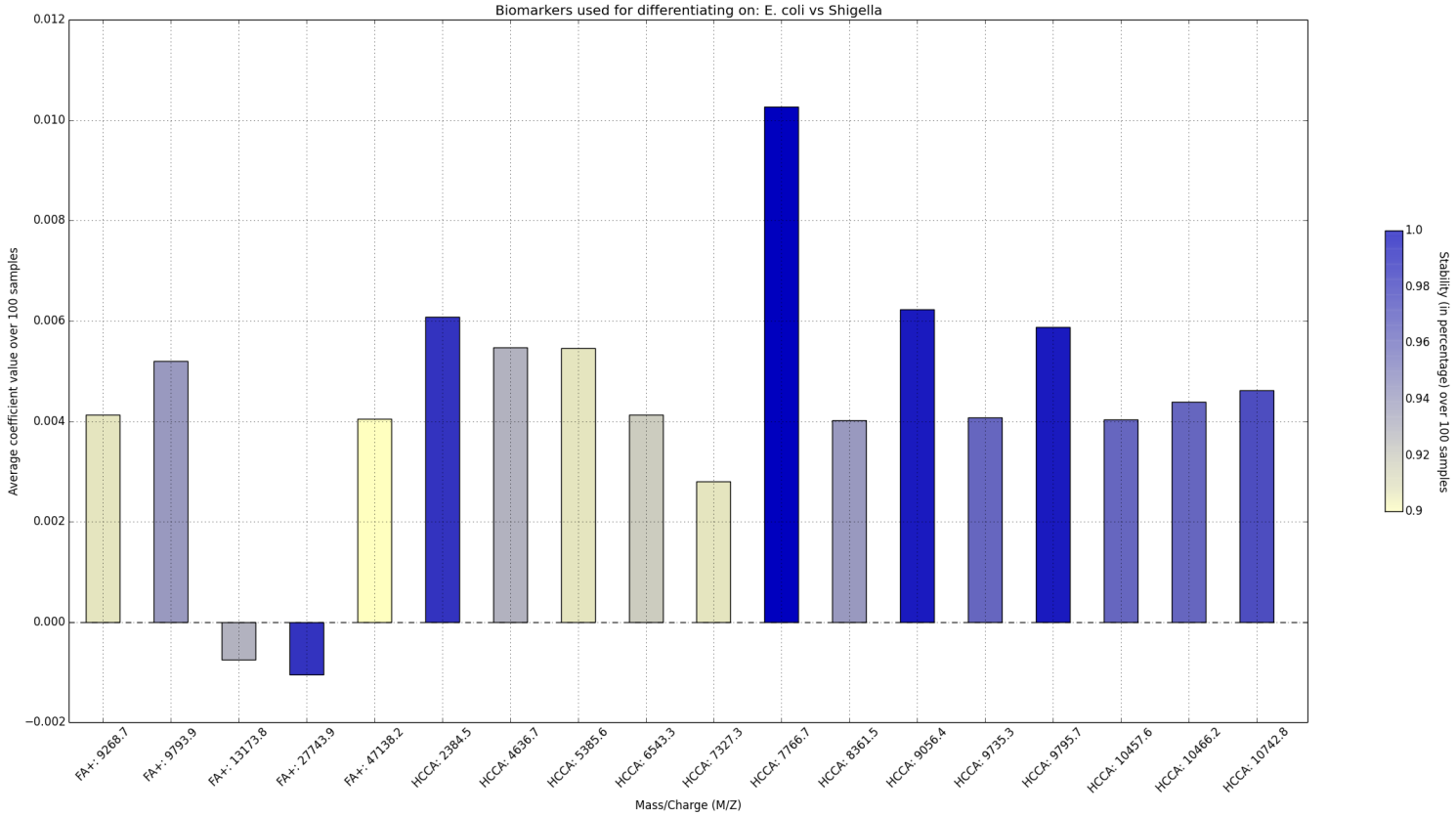

Figure A.4: The most stable features from the raw dataset.

References

Bennett Jr, W. E. & Tarr, P. I. (2009). Enteric infections and diagnostic testing. *Current opinion in gastroenterology*, *25*(1), 1–7.

Brenner, D. J., Fanning, G., Miklos, G. & Steigerwalt, A. (1973). Polynucleotide sequence relatedness among shigella species. *International Journal of Systematic Bacteriology*, *23*(1), 1–7.

Cannataro, M., Guzzi, P., Mazza, T. V. & Veltri, P. (2005). Preprocessing, management, and analysis of mass spectrometry proteomics data. *Workflows Management: New Abilities for the Biological Information Overflow–NETTAB*.

Cheng, A. C., McDonald, J. R. & Thielman, N. M. (2005). Infectious diarrhea in developed and developing countries. *Journal of clinical gastroenterology*, *39*(9), 757–773.

Coombes, K. R., Baggerly, K. A. & Morris, J. S. (2007). Pre-processing mass spectrometry data. In *Fundamentals of data mining in genomics and proteomics* (pp. 79–102). Springer.

Cortes, C. & Mohri, M. (2004). Auc optimization vs. error rate minimization. *Advances in neural information processing systems*, *16*(16), 313–320.

Deshpande, S., Jabbour, R., Snyder, P., Stanford, M., Wick, C. et al. (2011). ABOid: A software for automated identification and phyloproteomics classification of tandem mass spectrometric data. *J Chromatograph Separat Techniq S*, *5*, 2.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, *7*(3), 397–416.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, *3*, 1157–1182.

He, Y., Li, H., Lu, X., Stratton, C. W. & Tang, Y.-W. (2010). Mass spectrometry biotyper system identifies enteric bacterial pathogens directly from colonies grown on selective stool culture media. *Journal of clinical microbiology*, *48*(11), 3888–3892.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hsieh, S.-Y., Tseng, C.-L., Lee, Y.-S., Kuo, A.-J., Sun, C.-F., Lin, Y.-H. & Chen, J.-K. (2008). Highly efficient classification and identification of human pathogenic bacteria by MALDI-TOF MS. *Molecular & cellular proteomics*, *7*(2), 448–456.

Huang, J. & Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, *17*(3), 299–310.

Jaureguy, F., Landraud, L., Passet, V., Diancourt, L., Frapy, E., Guigon, G., ... others (2008). Phylogenetic and genomic diversity of human bacteremic escherichia coli strains. *BMC genomics*, *9*(1), 560.

Khot, P. D. & Fisher, M. A. (2013). Novel approach for differentiating shigella species and escherichia coli by matrix-assisted laser desorption ionization–time of flight mass spectrometry. *Journal of clinical microbiology*, *51*(11), 3711–3716.

Meinshausen, N. & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–

473.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Qiu, X., Gao, W. & Huang, X. (2009). Hierarchical multi-class text categorization with global margin maximization. In *Proceedings of the acl-ijcnlp 2009 conference short papers* (pp. 165–168).

Rifkin, R. & Klautau, A. (2004). In defense of one-vs-all classification. *The Journal of Machine Learning Research*, *5*, 101–141.

Saeys, Y., Inza, I. & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, *23*(19), 2507–2517.

Schmidt, H. & Jirstrand, M. (2006). Systems biology toolbox for matlab: a computational platform for research in systems biology. *Bioinformatics*, *22*(4), 514–515.

Secker, A., Davies, M. N., Freitas, A. A., Timmis, J., Mendao, M. & Flower, D. R. (2007). An experimental comparison of classification algorithms for hierarchical prediction of protein function. *Expert Update (Magazine of the British Computer Society's Specialist Group on AI)*, *9*(3), 17–22.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tong, D. L., Boocock, D. J., Coveney, C., Saif, J., Gomez, S. G., Querol, S., ... Ball, G. R. (2011). A simpler method of preprocessing MALDI-TOF MS data for differential biomarker analysis: stem cell and melanoma cancer studies. *Clinical proteomics*, *8*(1), 14.

Wagner, M., Naik, D. & Pothen, A. (2003). Protocols for disease classification from mass spectrometry data. *Proteomics*, *3*(9), 1692–1698.

Yang, C., He, Z. & Yu, W. (2009). Comparison of public peak detection algorithms for maldi mass spectrometry data analysis. *BMC bioinformatics*, *10*(1), 4.

Zhang, F. Z. & Hong, D. (2011). Elastic net-based framework for imaging mass spectrometry data biomarker selection and classification. *Statistics in medicine*, *30*(7), 753–768.

Zhang, X., Zhu, S., Xiong, Y., Deng, C. & Zhang, X. (2013). Development of a maldi-tof ms strategy for the high-throughput analysis of biomarkers: On-target aptamer immobilization and laser-accelerated proteolysis. *Angewandte Chemie International Edition*.

Zhang, Y. & Lin, K. (2012). A phylogenomic analysis of escherichia coli/shigella group: implications of genomic features associated with pathogenicity and ecological adaptation. *BMC Evolutionary Biology*, *12*(1), 174.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.