Queries for apls-66-01-30

This manuscript/text has been typeset from the submitted material. Please check this proof carefully to make sure there have been no font conversion errors or inadvertent formatting errors. Allen Press.

Bayesian Probability Approach to Feature Significance for Infrared Spectra of Bacteria

LUTZ HAMEL and CHRIS W. BROWN*

Department of Computer Science and Statistics (L.H.), and Department of Chemistry (C.W.B.), University of Rhode Island, Kingston, Rhode Island 02874

The significance of a spectral feature is defined as the probability that the feature captures the structure of the data set at hand. In particular, the significance is equal to a value proportional to the variance of a feature within a particular data set. The larger the variance, the higher the probability that the feature will capture the underlying structure. This approach is particularly useful when significance is used to select features differentiating clusters of samples and for the construction of self-organizing maps (SOMs) of clusters. A significance spectrum is obtained by plotting significance as a function of wavenumber. After developing the approach for feature significance, the significance framework was applied to the construction of SOMs for clustering infrared spectra of bacteria. The significance framework consistently chooses features that make it possible to construct maps with reduced feature sets that are at least as good as the maps constructed on full feature sets. In addition, significance reliably picks features that are consistent with biological interpretations of the spectra.

Index Headings: **Infrared spectra; Spectroscopy; Feature selection; Self-organizing maps; Significance spectrum; Bacteria spectra**.

INTRODUCTION

We have been exploring various methods for differentiating bacteria from their mid-infrared spectra. These spectra contain a large amount of information, allowing us to make inferences on chemical composition and structure of bacteria in addition to identification. For example, through spectroscopic analysis it is possible to differentiate the same class of bacteria based on the kind of agar substrate on which the culture was grown. Another example is the differentiation of the vegetative versus spore state of bacteria. Each state has a specific spectroscopic "signature" due to chemical differences.[1] For example, spores contain a peptidoglycan that is less cross-linked than in the vegetative cell. They also contain dipicolinic acid that is not found in the vegetative cells.[2]

The threat of bio-terrorism has intensified the development of rapid detection methods to monitor the purity of air, food, and water supplies. For the last two decades, Naumann and colleagues[3–16] have promoted mid-infrared spectroscopy as a rapid method of identifying microbes. In addition to their work, a number of other extensive investigations have been reported in the literature.[17–44] It is well known that the major features in the IR spectra of any biological system are primarily those of proteins. Thus, the spectral patterns of different genius-species of bacteria are very similar and we must rely on minor differences to identify a sample. As a consequence, in addition to measuring spectra of different genius-species, extensive efforts have been devoted to developing statistical methods for processing the spectra.[45–50] The statistical methods have included but are not limited to multivariate analysis (MVA), principal component analysis (PCA), artificial neural networks (ANN), partial least squares (PLS), discriminate analysis, derivatives, and peak heights.[50] Herein, we take a different approach to understanding the contributions that the spectral features make to the various statistical methods and apply this understanding to forming self-organizing maps for clustering different bacteria.

The Kohonen[51] neural network method was used to generate self-organizing maps (SOMs), which are a form of unsupervised learning.[52] Lavine et al.[53] have previously explored the use of SOMs for pattern recognition of infrared spectra. However, the feature selection discussed there is radically different from the probabilistic feature selection discussed here. Two recent papers have applied self-organizing maps to the problem of bacterial identification;[17,54] however, their approach to feature selection is also quite different from ours. For the most part, these methods consist of principal component analysis, in which it is difficult to obtain feature significance results over the whole spectrum of a particular data set.

From a data analysis point of view, spectra are not easy to work with due to their high dimensionality. A typical spectrum can have several thousand features, where each feature represents absorption at a particular wavelength. Herein, we propose a Bayesian approach to computing the significance of features in the absorption spectra. The advantage of this approach is that it is straightforward to compute and that it provides a probabilistic framework for feature selection over the whole spectrum. We are particularly interested in feature significance for two reasons: (1) as an aid in the interpretation of spectroscopic data, and (2) as a way to perform feature selection.

The ability to eliminate insignificant features from data sets with several thousand features can speed up training tremendously. Furthermore, insignificant features typically are noisy and can distort results if not eliminated. Herein, the methodologies for the Bayesian significance framework and self-organizing maps are discussed, validated, and then applied to sets of bacteria spectra.

THEORY

Feature Significance. The key insight of the present approach to feature significance is that the larger the variance of a feature in a data set the more likely it is to contribute to the clustering or grouping of the data set (an insight that also lays the foundation for principal component analysis). That is, the larger the variance of a feature, the more likely it is to capture the available structure in a data set. Consider Fig. 1; this is a scatter plot of a hypothetical, two-dimensional data set. Feature x1 has a large variance and captures the structure of the clusters

Received 6 October 2010; accepted 21 October 2011.

* Author to whom correspondence should be sent. E-mail: cbrown@chm.uri.edu.

DOI: 10.1366/10-06155

0003-7028/12/6601-000000$2.00/0

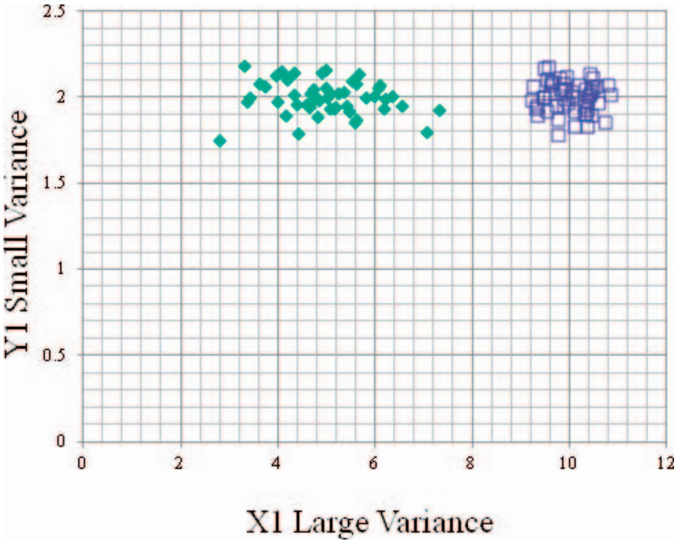

FIG. 1. Scatter plot of a hypothetical two-dimensional data set.

within the data. On the other hand, feature y1 has a small variance and does not capture any discernable structure within the data. Now, this case is somewhat idealized. Usually, having a feature with large variance does not necessarily mean that it encodes the available structure, but it means that the *probability* that it encodes the available structure is higher than that of features with small variance. Thus, in this case feature x1 is *significant*, since it has a higher probability of capturing the structure of the data than feature y1.

Bayesian Definition of Significance. The Bayesian definition of feature significance is based on the notion that a feature with large variance has a higher probability of capturing the available structure in a data set than a feature with small variance. We use Bayes' theorem to turn observed variances, also called observed significances, into the probability that a feature is significant. In order to do that we define the following quantities:

$$P(A_i|\ +) \equiv \text{observed significance of feature}\, A_i$$

$$P(+|\ A_i) \equiv \text{probability that feature}\, A_i \text{ is significant (significance)}$$

$$P(A_i) \equiv \text{prior probability of feature}\, A_i$$

Inserting these definitions into Bayes' theorem gives the following relation:

$$P(+|A_k) = \frac{P(A_k|+)P(A_k)}{\sum_{\text{i}} P(A_i|+)P(A_i)} \tag{1}$$

By assuming constant prior probabilities for all features, the equation can be simplified:

$$P(+|A_k) = \frac{P(A_k|+)}{\sum_{\text{i}} P(A_i|+)} \tag{2}$$

The probability that feature A_k is significant is computed by dividing its observed variance by the sum of the observed variances of all features in the data set.

As an example, we apply the Bayesian probability to the data set shown in Fig. 1. The observed variance of x1 is $P(\text{x1}|+) = 7.55$ and the observed variance of y1 is $P(\text{y1}|+) = 0.26$. Therefore, the sum of the observed variances is 7.81. Assuming constant prior probabilities for x1 and y1, we compute the significances as $P(+|\text{x1}) = 7.55/7.81 = 0.97$ and $P(+|\text{y1}) = 0.26/7.81 = 0.03$. This means that x1 has a 0.97 probability of capturing the available structure in the data compared to a probability of 0.03 that feature y1 captures the available structure.

The notion of feature significance defined as the probability that a feature encodes available structure in a data set gives a probabilistic framework for feature selection that is particularly well suited for self-organizing maps. The key to this probabilistic feature selection framework is the realization that the area under the significance curve is a probability mass, i.e., summing over the area under the curve for all features will give us a probability of 1. Rephrasing this slightly, the probability that the features encoded the available structure in the data set is equal to 1 if we use the entire feature set.

Self-Organizing Maps. SOMs are an unsupervised method for clustering or categorizing data.[52] Outliers have very little effect on the results of this method and it is not necessary to know the relationship between spectra and their categories. The method uses the Kohonen[51] neural network to produce SOMs of spectral data. This is a competitive learning algorithm in which a two-dimensional map is generated by neurons competing for each input spectrum. Given an input spectrum, the most similar neuron on the map is considered the winner, and it and its neighboring neurons are adjusted to have features similar to the input spectrum. Now, given another input spectrum the map location is found in the same way, i.e., the closest neuron to the spectrum is the winner, and it along with its neighbors is adjusted to match the inputted spectrum. This processing is continued for each of the input spectra and repeated until the map is converged. The effect of this processing is to produce a map of clusters in which each neuron at a specific location on the grid looks like a real spectrum. In some sense we could envision that the self-organizing map algorithm samples the underlying spectrum space. The neurons that represent clusters in the data will look very similar to actual spectra in its vicinity. Neurons that represent clusters in the data are often referred to as centroids.

The process of forming a SOM for infrared spectra is initiated by deciding upon the size of the map grid. The grid is two dimensional and can be either square or rectangular in shape. An example 8×8 square grid of neurons is shown in Fig. 2a; each of the neurons is represented by a random number spectrum, which has the same length as the spectra to be classified. The size of the map can be fine-tuned using specific convergence criteria. The random number spectra are generated using values between 0 and 1 at each of the wavenumbers in the spectrum. A different random number spectrum is placed at each centroid.

To demonstrate the development of a SOM, we use five library spectra of simple organic liquids shown in Fig. 2b. The processing starts by inputting the first library spectrum on the right. Its Euclidean distance from each of the 64 random number spectra is calculated and the winner is the neuron

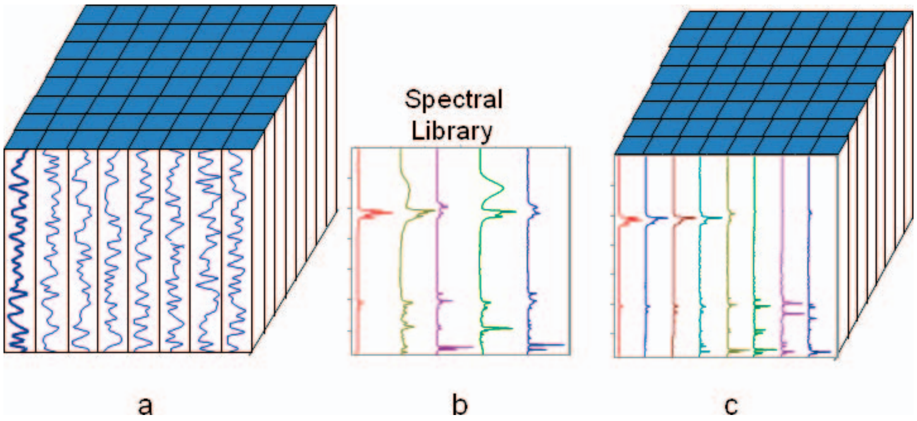

FIG. 2. (**a**) Grid of neurons as random number spectra. (**b**) Library spectra for training neurons. (**c**) Grid of neurons trained with library spectra.

having the shortest distance, i.e., the random number spectrum most similar to the library spectrum. The random number spectrum at the winning neuron is adjusted to be closer to the input spectrum. In addition, all neighboring neurons are adjusted to be similar to the winner using a bubble neighborhood function,[51] which is a constant function in the defined neighborhood of the winner neuron. On the first pass through all of the spectra, the entire map of neurons is considered as neighbors and all are adjusted; the size of the neighborhood of the winner neuron decreases with each iteration through the input data set by decreasing the width of the bubble neighbor.

A second spectrum from the five-spectrum library is inputted to the network, its Euclidean distance from each of the neurons is calculated, and the winning neuron is found. The map is adjusted so that the winning neuron is similar to the input spectrum and all of the other neurons on the map are modified to appear closer to the winner as described above. This process is repeated for each spectrum in the library; the winning neuron and its neighbors are adjusted to be closer to the inputted spectrum. Obviously, the entire map is constantly changing with the addition of each input spectrum since, during the first pass, all of the map neurons are adjusted with each inputted spectrum. After the first pass, the map has been adjusted n times, where n is the number of input spectra.

After processing all of the library spectra during the first pass, the entire procedure discussed above is repeated again on each of the input spectra one at a time with a smaller neighborhood. This processing is applied iteratively until some convergence criterion is reached. In our case, the convergence of the map is computed by considering both the neurons of the map and the input spectra as two distinct samples from the same underlying spectrum distribution. If the two samples appear to be drawn from the sample distribution under a two-sample test such as the F-test, then we say that the map is converged.[55] The neurons or centroid spectra in the final grid appear like real spectra. An example for the first row in the grid is shown in Fig. 2c. In this example, from the left, the first three centroid spectra are from the alkane region, the next two are from the alkyl aromatic region, and the last three are from the general aromatic region with halogen or nitro substitutions. More exact details of these various regions will be given for bacteria sets in the Results and Discussion section.

EXPERIMENTAL

Growth of Bacteria. The original bacteria samples were from the American Tissue Culture Collection (ATCC) and were cultured from in-house stocks. Cells of the vegetative form of *Bacillus cereus* were cultured on agar plates of chocolate blood, blood, nutrient, and mannitol to determine the effects of agars on the spectra. Vegetative cells of all genus-species were cultured on nutrient agar plates. With the exception of *Psuedomonas fluorescens,* which was grown at 30 °C, the cells were cultured at 37 °C for 48 hours. The spores were grown on nutrient agar plates at 30 °C for 3 to 5 days. The vegetative bacteria and spores were harvested from the plates in distilled water and pelleted by centrifugation (10 000 rpm in an SS34 rotor for 10 minutes). The pellets were re-suspended in distilled water and centrifuged two additional times. To collect the spores and to eliminate the vegetative bacteria, the samples were treated with lysozyme overnight and washed with distilled water and centrifuged two more times to remove vegetative bacterial cells while leaving the spores intact.

Instrumental. All spectra were measured using a TravelIR (Smiths Detection, formerly SensIR, Danbury, CT) Fourier transform infrared (FT-IR) spectrometer with a single-bounce ZnSe/diamond crystal attenuated total reflection (ATR) accessory. The spectra were obtained at a resolution of 4 cm^{-1} with 64 co-added scans over the spectral range of 4000–650 cm^{-1}. These scans were averaged to provide the final spectrum for each sample. ATR reference spectra were measured using the blank diamond crystal.

Data Pretreatment. Data pretreatment was executed using Matlab 7.1 (Mathworks, Natick, MA), including the Neural Network Toolbox. To remove baseline effects, the spectra were converted to first derivatives using a Savitzky–Golay 13-point cubic smoothing conversion. All spectra were normalized to a total area of 1.0.

Software for Data Analysis. All subsequent data analysis and self-organizing map construction was done in R (www.r-project.org) using custom extensions to the 'som' package. The custom extensions include an enhanced annotation of the unified distance matrix to make clusters more visible.[56] Briefly, this consists of tracing the gradient on the unified distance matrix to the "low points" where the gradient is zero. This tracing gives rise to star graphs with the "low points" at their centers and these star graphs correspond precisely to the underlying clusters. The custom extensions also include an implementation of the two sample convergence criteria explained briefly above.[57]

RESULTS AND DISCUSSION

The average spectra of a set of twenty *B. cereus* samples with each set grown on four different agars are shown in Fig. 3. The first-derivative spectra are compared in the middle of the figure and the resulting significance spectrum obtained from the first-derivatives is shown at the bottom. Contributions to the significance spectrum are more apparent from the first-derivative spectra than from the original spectra. The contributions to the significance at ~1700 cm^{-1} are apparent in the original absorption spectrum; however, the other contributions are more apparent in the derivative spectra, although some of these are difficult to visualize.

Sorting features (wavenumbers) in decreasing order of significance provides a probability distribution for the feature set. This gives rise to a graph as shown in Fig. 4. Using this probability distribution, we can now determine the features that should be included in order to have confidence (e.g., 90%) that the available structure in the data set has been captured. This means that the area under the curve to the left of the line marked "p = 0.9" in Fig. 4 has a probability mass of 0.9. The features that contribute to this probability mass are the features that we would use for model building. Given this probabilistic feature selection framework, we can also ask questions such as the probability that the ten most significant features capture the structure of the data set. In order to answer this question, we simply add the significances of the ten most significant features.

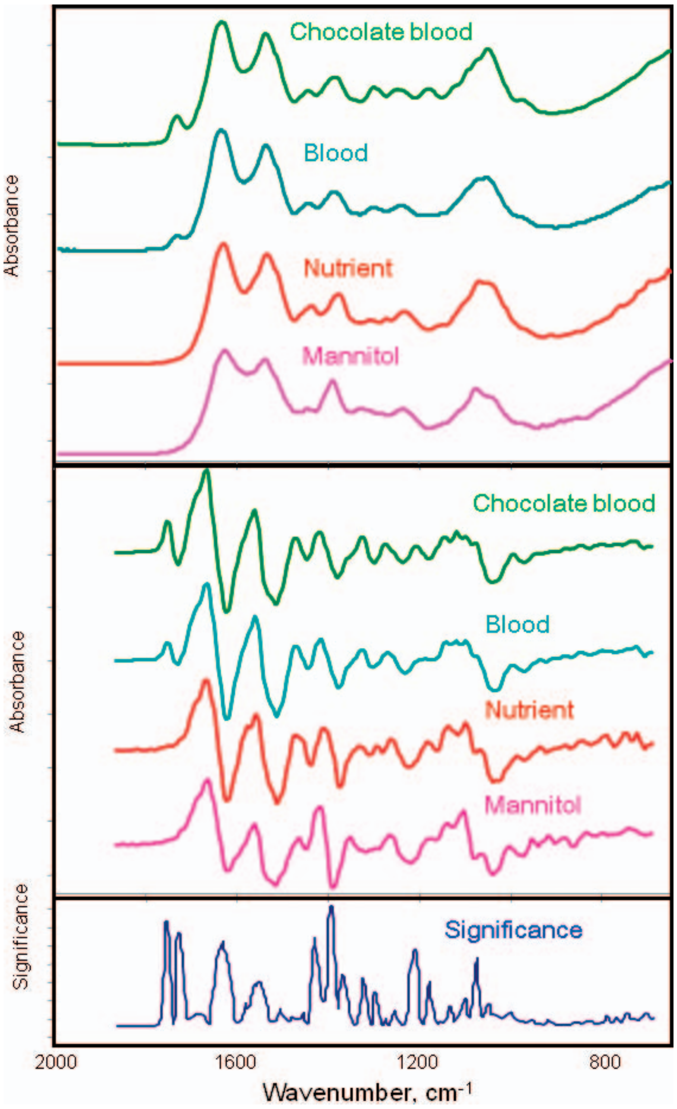

FIG. 3. Original spectra, first derivatives, and significance spectrum for *Bacillus cereus* grown on different agars.

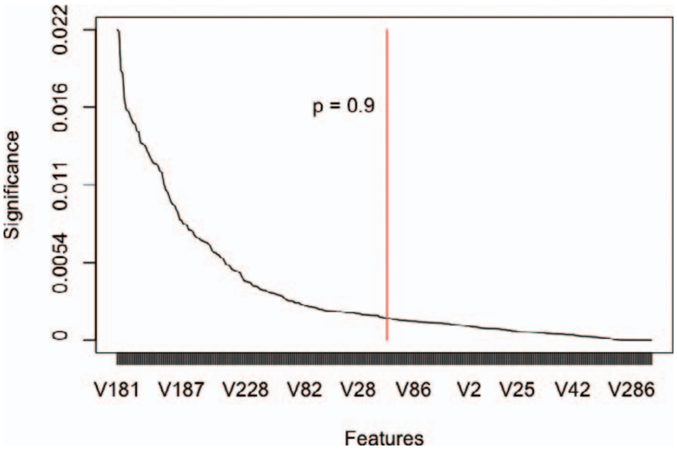

FIG. 4. Feature probability distribution. The area to the left of the 0.9 vertical line represents 90% of the total area.

Observe that here we make no attempt to remove highly correlated features: first, because we are interested in the significance of all features when evaluating the bacteria spectra, and second, we found during our experiments that highly correlated features did not pose a particular problem for constructing self-organizing maps.

Validation. A set of statistical experiments designed to validate our notion of a significance spectrum was performed. Inspired by the map stability as presented by Cottrell et al.,[57] the present experiments are also based on that stability. However, in this case, map stability is based on intra-cluster geometry with the idea that significant features will preserve intra-cluster geometry in repeated map building exercises and features that are not significant will not. Specifically, stability is defined as the inverse of the variance of the pairwise distances between the members of clusters. A Monte-Carlo estimate of the map stability is used to construct M randomly initialized maps and to compute the variability of the intra-cluster distances based on these Monte-Carlo samples. Each Monte-Carlo sample consists of a randomly initialized map with the dimensions xdim and ydim. Once initialized, the map is trained with the training data. We then measure the distances between the intra-cluster pairs on the map. With the M samples we compute variance in the distance between each pair. Finally, we compute the average variance over all the pairs and then return the inverse of the average variance as the stability.

In the first experiment, we investigated how effective the feature selection is with respect to separating significant features from features that are not significant. The underlying assumption is that significant features will maintain the overall geometry of the clusters across all the Monte-Carlo samples implying low variability and therefore high stability. On the other hand, features that are not significant typically do not

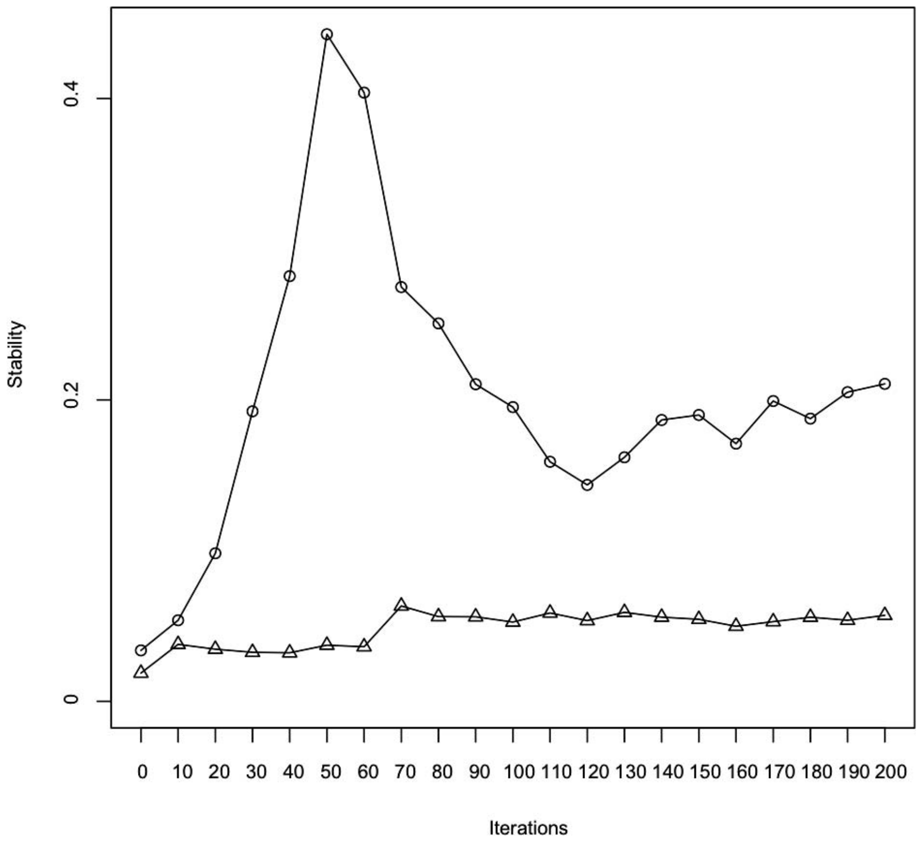

FIG. 5. Stability of the 15 most significant features (circles) versus stability of 15 non-significant features (triangles).

encode enough structure in order to maintain stable cluster geometries across the Monte-Carlo samples. In this experiment, we used a bacterial spectra data set with 300 features (absorbances at 300 wavenumbers). Two subsets were selected from the data set: one subset consisted of the 15 most significant features and the other subset consisted of the 15 least significant features. The stability test was applied to these data sets and the results are shown in Fig. 5. The *x*-axis is the number of training iterations we applied to the maps and the *y*-axis is the stability. The curve with the circles is the stability curve due to the 15 significant features and the curve with the triangles is the stability curve due to the 15 least significant features. As expected, the maps due to the significant features are much more stable with increasing number of training iterations than the maps due to the least significant features. This clearly shows that our feature selection separates significant features from features that are not significant.

Currently, we do not have an explanation for the bump in stability of the significant features at about 50 training iterations. We have observed this phenomenon in other bacterial spectra data sets and intend to investigate. Herein, we are only concerned with the difference in stability between significant and non-significant features.

Significant Features versus Full Feature Set. In this next experiment, the stability of maps built using a data set that consists only of the 15 most significant features were investigated. The map stability of this reduced data set was compared to the stability of maps constructed with the full feature set. Our expectation was that the maps constructed with the significant features should be as stable as the maps constructed with the full feature set. The stabilities of both kinds of maps are shown in Fig. 6. As before, the *x*-axis describes the number of training iterations applied to the maps and the *y*-axis describes the stability. The curve with the circles is due to the stability of the maps constructed using the reduced data set. The curve with the triangles is due to the stability of the maps constructed with the full data set. As can be seen, the curves are very similar to each other, validating our expectation. Furthermore, it seems that the maps due to the reduced data set are slightly more stable than the maps constructed with the full data set. We suspect that the slight instability in the maps constructed using the full feature set is due to non-significant features introducing noise.

Statistical Significance of our Feature Selection. Here we investigate the performance of our feature selection based on selecting 15 of the most significant features from the available feature set and comparing this to a feature selection based on randomly choosing 15 features from the available feature set. We use a Monte-Carlo estimate of the 90% confidence intervals. Training each of the maps with 200 iterations on the respective data sets gives us the results in Table I. Notice that the feature selection based on feature significance gives rise to maps that on average are more stable than maps based on random feature selection. However, the 90% confidence intervals touch (and will probably overlap if we were to consider 95% confidence intervals) meaning that statistically there is a chance that random feature selection will perform as well as our feature selection based on feature significance. This is perhaps not surprising, since as soon as the random feature selection picks out one of the more significant features from the feature set then it will perform well. However, also notice that the confidence band for the random feature selection is much

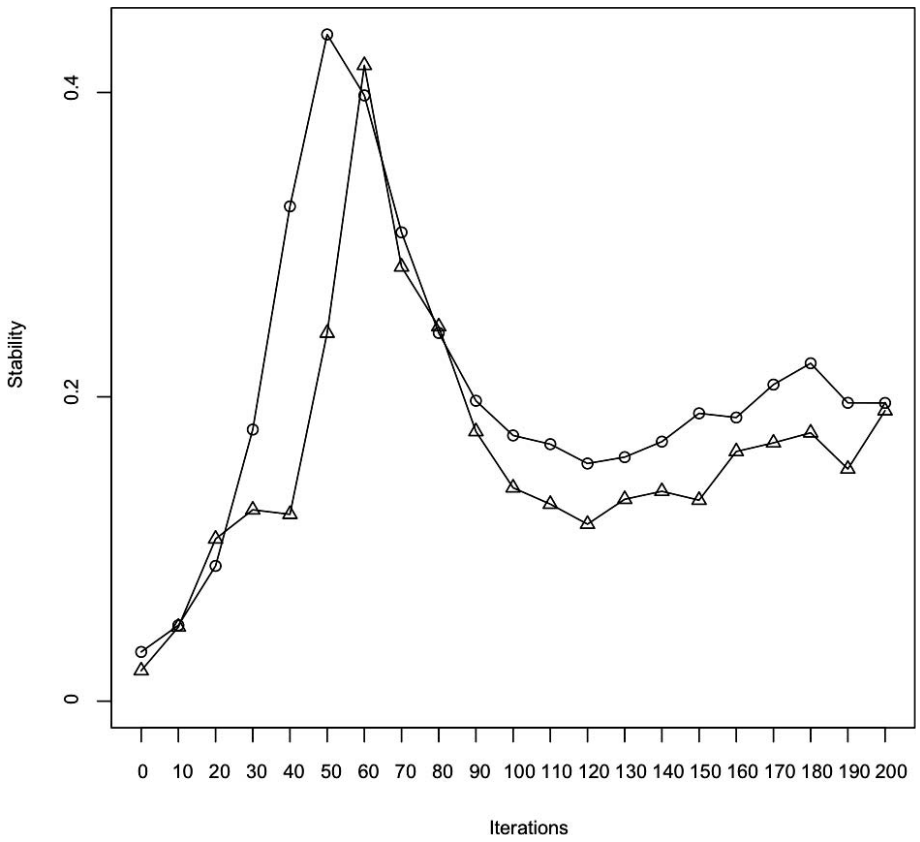

FIG. 6. Stability of the reduced set of 15 features (circles) versus the full feature (300) set (triangles).

wider than the confidence band for the significance-based feature selection, implying that there is a lot of variability in the performance of the random feature selection compared to the significance-based feature selection. This means that sometimes we can "get lucky" with random feature selection and sometimes not. The performance of the significance-based feature selection is much more consistent. We also compared our feature selection with PCA and found that both agree on the top 25 features and agree mostly on the top 100 features. The difference in the remaining features is the way each of the algorithms is biased; our feature selection interprets variance directly as a way to capture structure whereas PCA is biased by its linearity assumptions.

Bacteria Spectra. Three bacterial data sets are considered in this section. The spectra cover the range from 670 to 1870 cm^{-1}. The experiments were designed to highlight our significance framework in both feature selection and as an aid in the interpretation of the spectra.

Spores vs. Vegetative Cells. The first data set consists of 53 spectra of the spores and vegetative cells of two species of bacteria: *Bacillus thuringiensis* and *Bacillus subtilis*. A starburst unified distance matrix[56] for a self-organizing map constructed on the full feature set is shown in Fig. 7. In addition to the color-coding of the unified distances on the map, the starbursts highlight the clusters. The labels St and Vt label the spores and vegetative cells of *Bacillus thuringiensis* on the map, respectively, whereas the labels Ss and Vs label the spores and vegetative cells of *Bacillus subtilis* on the map, respectively. It is easy to see that spores and vegetative cells are fully separated; the spores appear on the left of the map and the vegetative cells appear on the right. Dark colors on the unified distance matrix indicate a strong separation. Here this means that there is a substantial spectral difference between spores and vegetative cells. Furthermore, we see that each species of bacteria forms it own cluster within the spores as well as within the vegetative cells. This leads us to believe that there are not only spectral differences between spores and vegetative cells but also spectral differences between the bacteria species in general.

TABLE I. Statistical significance of feature selection.

	Mean stability	90% Confidence interval
Significance selection	0.20	0.22–0.18
Random selection	0.12	0.18–0.08

The unified distance matrix for a map constructed with a reduced feature set is shown in Fig. 8. The feature selection picked 128 features out of 300 to construct the self-organizing map with a 90% probability of significance. A close examination of the map reveals that all the essentials are preserved in this map. Moreover, there is clear distinction between spores and vegetative cells and a clear demarcation between the individual species of bacteria. There is actually clearer clustering for the two species of the vegetative cells into two groups with this more limited data set compared to using all of the spectral features.

One of the main differentiating factors between spores and vegetative cells is the conformation of the peptidoglycan molecule. In spores it is less cross-linked compared to vegetative cells. Naumann et al.[3] have shown that these

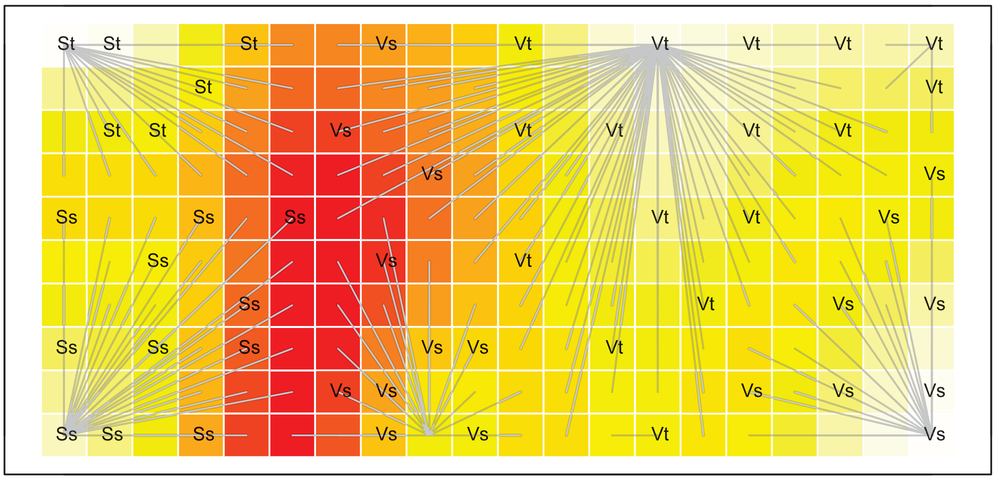

FIG. 7. SOM for spores (S) versus vegetative (V) cells, full feature set of 300 data points. Axes are relative *x* and *y* distances showing the map location of samples.

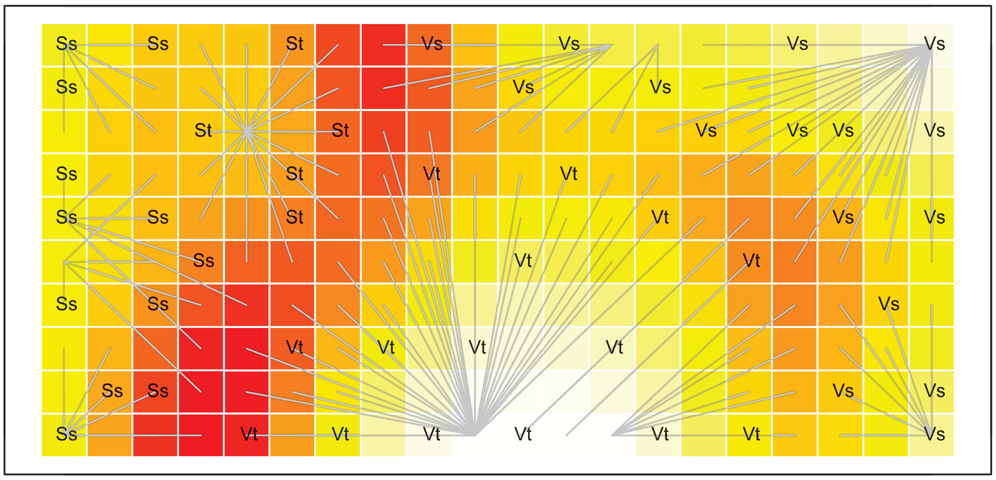

FIG. 8. SOM for spores (S) versus vegetative (V) cells, reduced feature set of 128 data points. Axes are relative *x* and *y* distances showing the map location of samples.

conformational differences of peptidoglycan produce drastic variations in absorbance near 1730 cm^{-1}. In our significance plot for this data set (shown in Fig. 9), two dominant bands are observed at ~1730 cm^{-1}. Postulating that these bands in the significance spectrum are due to the variation of absorbance stemming from the different conformations of the peptidoglycan molecule, we constructed a data set that just consisted of the two features highlighted by these two bands. The resulting self-organizing map can be seen in Fig. 10. It is remarkable that with just these two features we obtain a near perfect separation between spores and vegetative cells. Another interesting observation is that the identity of the individual species of bacteria is also preserved. We postulate that the conformations of the peptidoglycan are species specific. From the perspective of the Bayesian feature selection framework, we see that it assigned the band around 1730 cm^{-1} a high significance and this coincides with findings obtained independently (viz. Naumann et al.[3]).

Gram-Positive vs. Gram-Negative Bacteria. The next data set was concerned with the differentiation between Gram-positive (G(+)) and Gram-negative (G(-)) bacteria. A bacterium is G(+) if it retains the dark blue/purple color of the absorbed dyes due to Gram staining.[8] Bacteria that do not retain the dyes and therefore do not assume the typical bluish color are G(-). The data set used to investigate G(+) and G(-) bacteria consists of 340 observations, approximately half of which are G(+) and the other half are G(-). The G(+) bacteria include *Bacillus cereus* and *Bacillus subtilis* among others. The G(-) bacteria include *Escherichia coli* and *Salmonella typhimurium* among others. Overall the data set incorporates 17 different species of bacteria.

A self-organizing map constructed for the G(+) and G(-) data set is shown in Fig. 11. It is clear that there are spectral differences between G(+) and G(-) bacteria because they are clearly separated on the map. G(+) bacteria are labeled P and appear on the top left of the map and G(-) bacteria are labeled N and appear at the bottom right of the map. This map was constructed with a reduced feature set wherein the features were selected according to significance. The overall probability of significance adds up to 0.9. In this case, we used 153 of 300 available features to construct the map.

We investigated the difference between G(+) and G(-) bacteria further. In particular we were interested in whether our significant features had any biological relevance. Looking at the significance plot, Fig. 12, we see three prominent bands: two centered around 1730 cm^{-1} and one at 1200 cm^{-1}. It is no

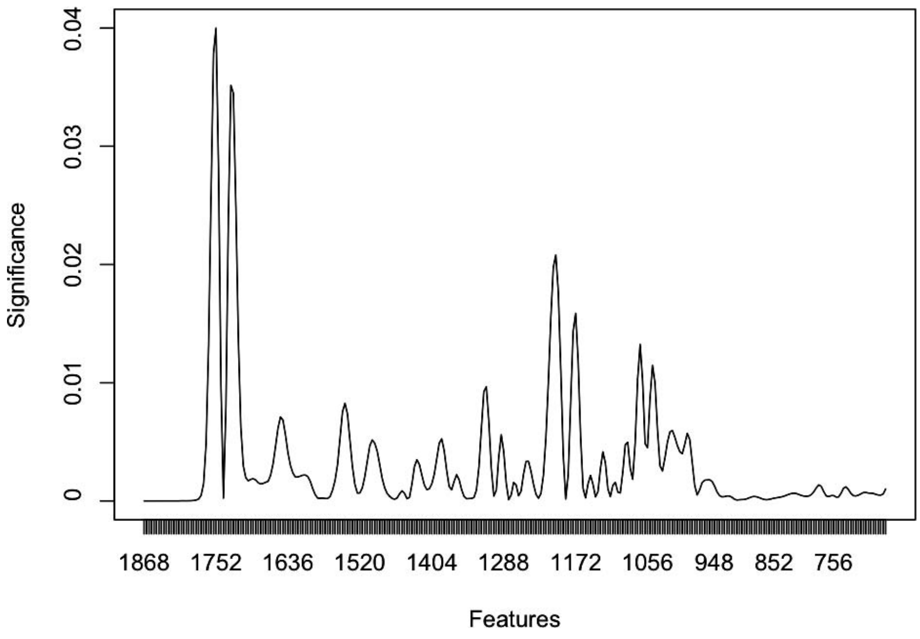

FIG. 9. Spores vs vegetative cells, significance plot. Features are in cm^{-1} units.

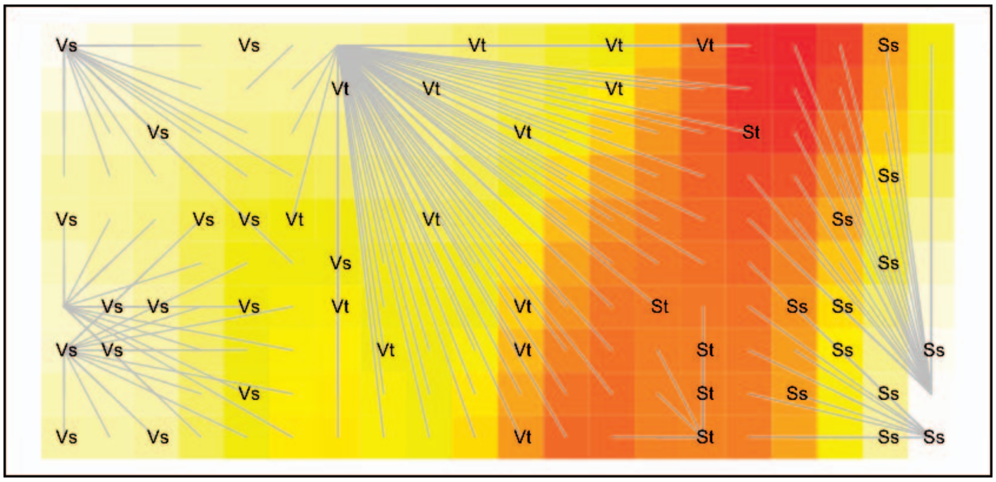

FIG. 10. SOM for spores (S) versus vegetative (V) cells, two features. Axes are relative *x* and *y* distances showing the map location of samples.

surprise that we see the bands at 1730 cm^{-1} again. Peptidoglycan plays a central role in the differentiation between G(+) and G(-) bacteria. G(+) bacteria possess a thick outer layer of peptidoglycan, enabling them to absorb and retain the Gram-stain crystals. G(-) bacteria, on the other hand, only possess a thin, internal layer of peptidoglycan that does not allow them to retain the staining crystals. This structural variation is expressed in the spectra and we see the two characteristic significance peaks around 1730 cm^{-1}. According to Naumann et al.,[3] the significance band at 1200 cm^{-1} is due to complex sugar ring modes. Unfortunately, there are no simple and generally accepted correlations between the three-dimensional arrangement of the sugar residues and the magnitudes of the absorptions in this region of the spectrum. However, we can use the infrared characteristics in this region as "fingerprints." A SOM constructed only from the three features indicated by the significance features is shown in Fig. 13. The G(-) bacteria occupy the top left part of the map and the remainder of the map is dedicated to the G(+) bacteria.

Effects of Culture Agars. In this experiment, the vegetative form of *Bacillus cereus* was grown on different agars: nutrient (N), mannitol (M), blood (B), and blood-chocolate (C). The letters in parentheses are the labels used on the map. The data set contains 88 spectra fairly evenly distributed over the four different agars with the exception of mannitol, for which there are only five spectra. A self-organizing map constructed using a reduced feature set at the 90% significance level is shown in Fig. 14. The bacteria grown on the four different agars are clearly separated, with mannitol at the top left, nutrient at the bottom right, blood at the top right, and blood-chocolate at the bottom left. This means that there are noticeable spectral differences between the bacteria grown on these four different agars as we mentioned earlier with respect to the spectra shown in Fig. 3.

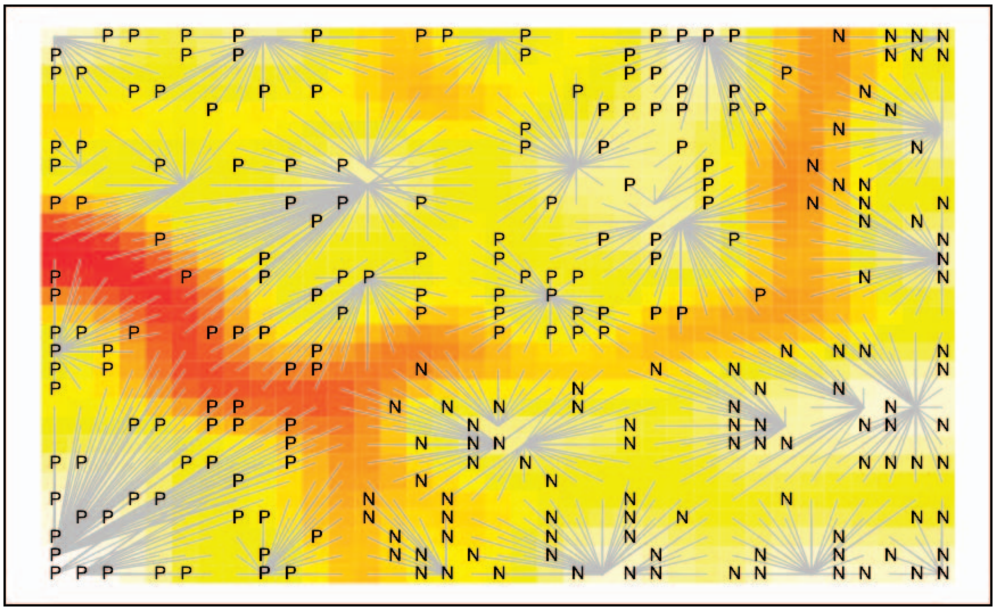

Fig. 11. SOM for Gram-positive (P) vesrus Gram-negative (N), reduced feature set, 90% significance. Axes are relative x and y distances showing the map location of samples.

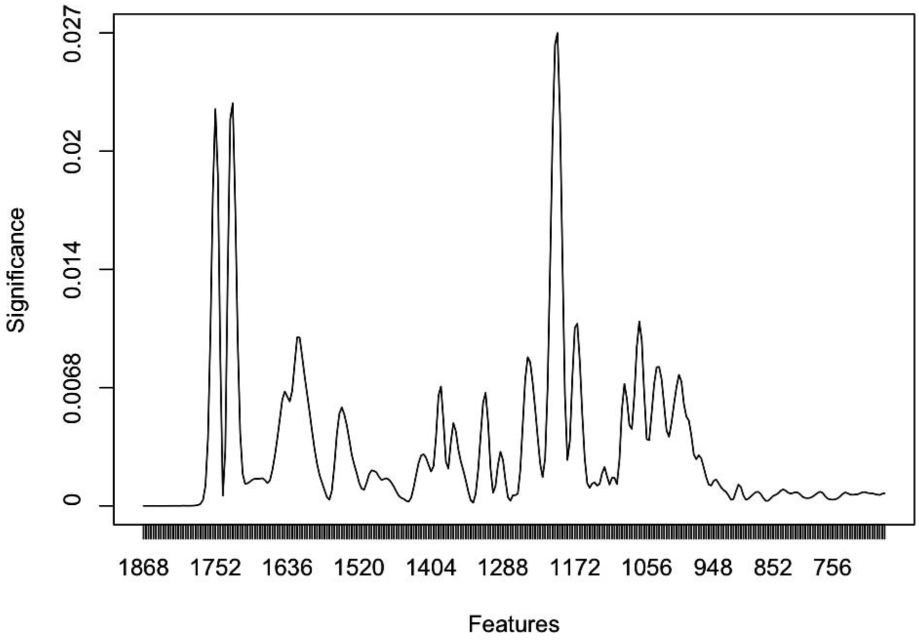

Fig. 12. Significance plot of Gram-positive versus Gram-negative. Features are in cm^{-1} units.

The significance spectrum for the four agars is shown in Fig. 15. The significance bands appear at ~1750, 1636, 1401, 1196, and 1064 cm^{-1}. The bands between 1200 and 1000 cm^{-1} are due to complex sugar ring modes[7] and we postulate that a large variation in these bands from one agar to another most likely indicates that the sugars from the different agars are metabolized in different ways. What was surprising in this analysis was the fact that large variations in bands typically associated with conformational differences in peptidoglycan (bands 1750, 1636, and 1401 cm^{-1}). These five bands were extremely predictive in terms of the underlying agars. Figure 16 shows a SOM constructed using only the five features due to the largest bands in the significance plot. As can be seen from the map, the bacteria grown on the four different agars were completely separated.

Comparisons with Other Statistical Methods. The present method was designed to highlight the most significant spectral features and use these to cluster samples from different categories on SOMs. Analysis of variance (ANOVA) has been used previously for feature selection by determining features that correlate with the classification of samples.[58,59] ANOVA has also been used in conjunction with principal component analysis (PCA) to determine biomarkers for premature births from MALDI-MS of amniotic fluids.[60] The ANOVA-PCA

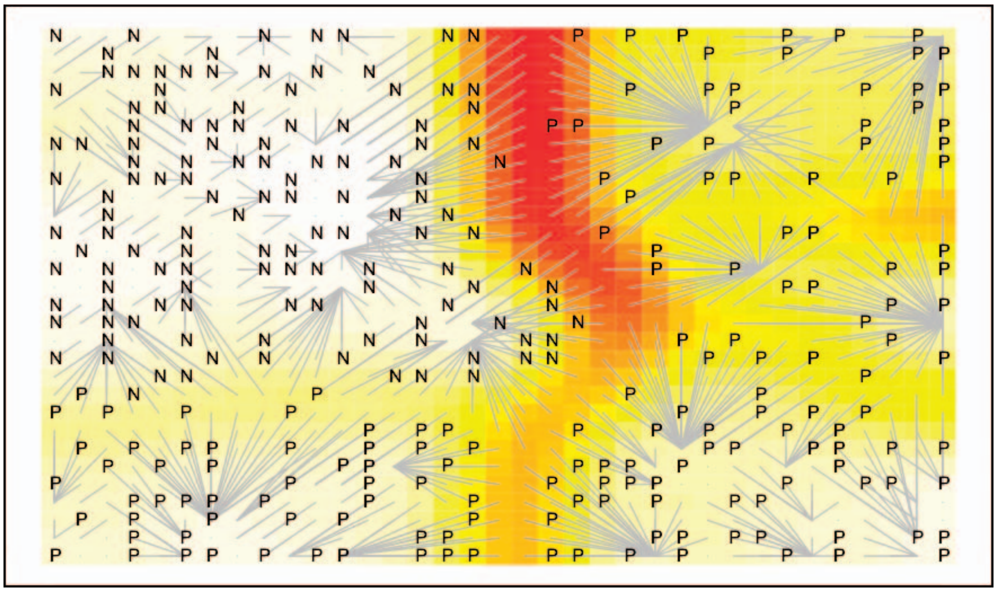

FIG. 13. SOM for Gram-positive (P) versus Gram-negative (N) with a reduced feature set of three features. Axes are relative *x* and *y* distances showing the map location of samples.

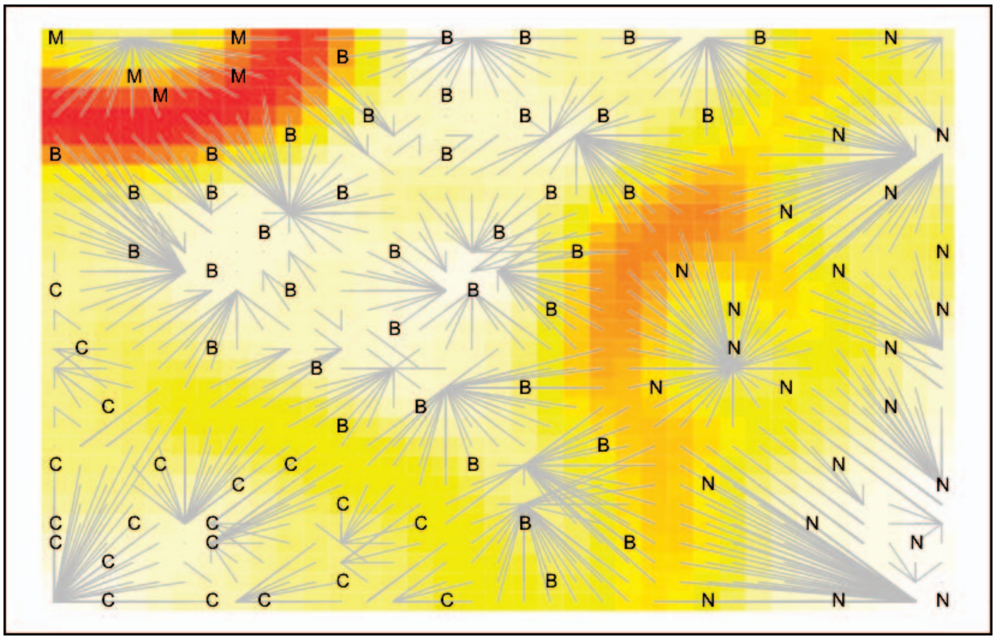

FIG. 14. SOM for *Bacillus cereus* grown on different agars, reduced feature set, 90%. M = Manitol, B = blood, C = chocolate blood, and N = nutrient. Axes are relative *x* and *y* distances showing the map location of samples.

combination has also been used to determine the effects of concentration and temperature on adhesion of carrageenan gels.[61] All of these investigations require training with samples from known classes to provide prediction models; the training is accomplished with labeled samples. The method demonstrated herein does not require that the class or type of samples be known prior to producing the significance spectra and the SOMs.

Other methods such as K-means[62] and hierarchical cluster analysis[63] (HCA) can be used to cluster samples from spectroscopic data. These approaches assign observations to clusters exclusively, i.e., an element can only belong to a single cluster. The advantage of SOM is that it can portray trends, such that in cases where cluster membership is somewhat ambiguous the algorithm will place that observation on the intersection between the clusters in question. Also, K-means has no visualization and the visualization for HCA is easily overwhelmed by just tens of points, never mind hundreds or thousands of points.

CONCLUSIONS

Significance of features in infrared spectra of bacteria has been defined as the probability that a feature captures structure

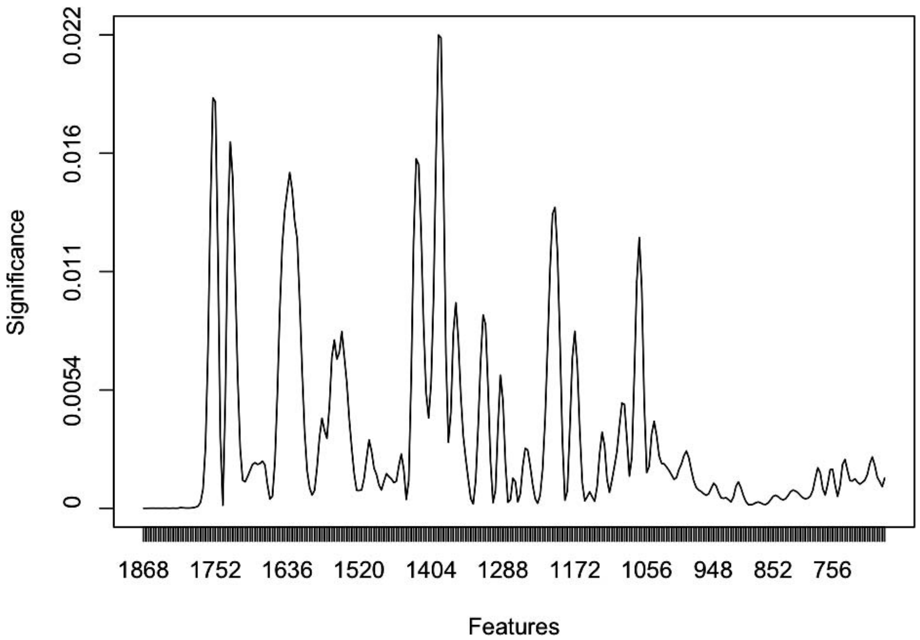

FIG. 15. Significance plot of *Bacillus cereus* grown on different agars. Features are in cm^{-1} units.

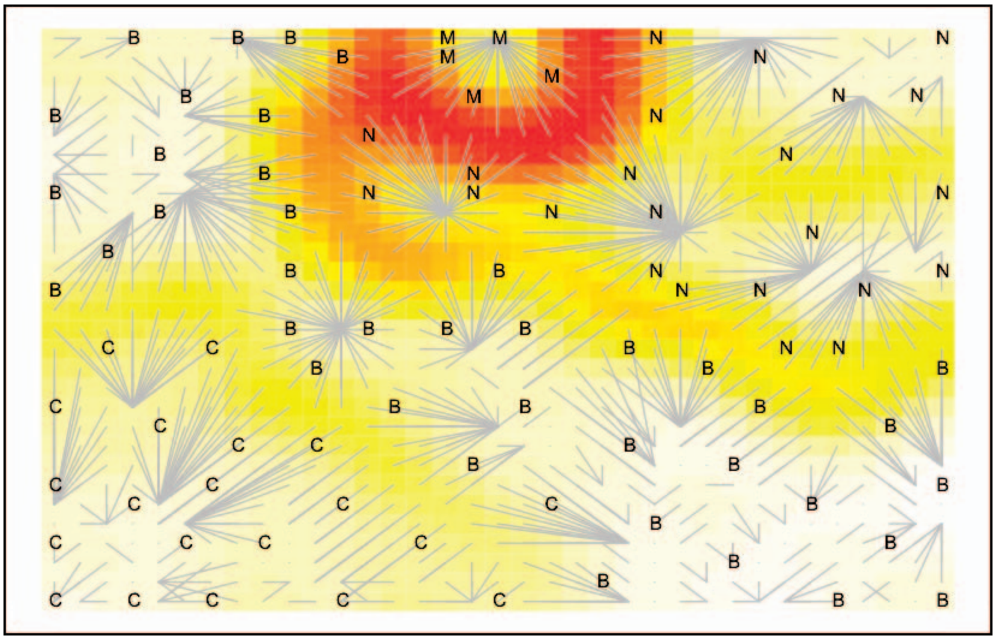

FIG. 16. SOM for *Bacillus cereus* grown on different agars with reduced set of five features. M = Manitol, B = blood, C = chocolate blood, and N = nutrient. Axes are relative *x* and *y* distances showing the map location of samples.

available in the data set. As such, significance of a feature is defined as a value proportional to the variance of this feature within a given library of such spectra. It was shown that this notion of significance seems to be appropriate for the analysis of bacterial spectra using self-organizing maps. Significant features lead to more stable maps than non-significant features. Using collections of real-world bacterial spectra, we have shown that our feature selection consistently chooses features that allow us to construct maps on reduced feature sets that are at least as good as the maps constructed on the full feature sets. In addition, our notion of significance reliably picks features that are consistent with the biological interpretation of the spectra.

We are interested in further exploring our notion of significance as an aid to the interpretation of bacterial spectra. One particular area we are interested in is the further understanding of the role of peptidoglycan in the differentiation of bacterial cells. The surprising effect of agars on the bacterial chemistry is also a worthwhile topic for investigation.

1. D. Helm, H. Labischinski, G. Schallehn, and D. Naumann, Microbiology, **137,** 69 (1991).

2. "Spore Structure and Chemistry, Vegetative Cells, Bacillus Spore, Dipicolinic Acid," http://www.microbiologyprocedure.com/microorganisms/spore-structure-and-chemistry.htm.
3. D. Naumann, G. Barnickel, H. Bradaczek, H. Labischinski, and P. Giesbrecht, Eur. J. Biochem. **125,** 505 (1982).
4. P. Giesbrecht, D. Naumann, H. Labischinski, and G. Barnickel, "A new method for the rapid identification and differentiation of pathogenic microorganisms using Fourier transform infrared spectroscopy," in *Rapid Methods and Automation in Microbiology and Immunology,* K.-O. Habermehi, Ed. (Springer, Berlin, 1985), p. 198.
5. D. Naumann, V. Fijala, H. Labischinski, and P. Giesbrecht, J. Mol. Struct. **174,** 165 (1988).
6. D. Naumann, D. Helm, and H. Labischinski, Nature (London) **351**, 81 (1991).
7. D. Naumann, D. Helm, H. Labischinski, and P. Giesbrecht, "The Characterization of Microorganisms by Fourier-Transform Infrared Spectroscopy (FT-JR)," in *Modern Techniques for Rapid Microbiological Analysis,* W. H. Nelson, Ed. (VCH Publishers, New York, 1991), p. 43.
8. D. Helm, H. Labischinski, and D. Naumann, J. Microbiol. Methods **14,** 127 (1991).
9. D. Helm, H. Labischinski, G. Schallehn, and D. Naumann, J. Gen. Miciobiol. **137**, 69 (1991).
10. D. Naumann, S. Keller, D. Helm, C. Schultz, and B. Schrader, J. Mol. Struct. **347,** 399 (1995).
11. D. Helm and D. Naumann, FEMS Miciobiol. Lett. **126**, 75 (1995).
12. H. C. van der Mei, D. Naumann, and H. J. Busscher, Infrared Phys. Technol. **37,** 561 (1996).
13. D. Naumann, D. Helm, and C. Schultz, "Characterization and Identification of Micro-Organisms by FT-IR Spectroscopy and FT-IFt Microscopy," in *Bacterial Diversity and Systematics,* F. G. Priest, A. R. Cormenzana, and B. J. Tindall, Eds. (Plenum Press, New York, 1994), pp. 67–85.
14. D. Naumann, "Infrared spectroscopy in microbiology", in R. A. Meyers, Ed., *Encyclopedia of Analytical Chemistry* (John Wiley and Sons, New York, 2000), vol. 1, pp. 102–131.
15. D. Naumann, "FT-IR and FT-Raman Spectroscopy in Biomedical Research," in *Infrared and Raman Spectroscopy of Biological Materials,* H.-U. Gremlich and B. Yan, Eds. (Marcel Dekker, New York, 2001), Chap. 9, pp. 323–378.
16. D. Naumann, "Infrared Spectroscopy in Microbiology," in *Encyclopedia of Analytical Chemistry: Applications, Theory, and Instrumentation; Biomedical Spectroscopy,* R. A. Meyers, Ed. (John Wiley and Sons, Chichester, 2000), vol. I, pp. 102–131.
17. N. M. Amiali, M. R. Mulvey, J. Sedman, M. Louie, A. E. Simor, and A. A. Ismail, J. Microbiol. Methods **68,** 236 (2007).
18. R. Goodacre, E. M. Timmins, P. J. Rooney, J. J. Rowland, and D. B. Kell, FEMS Microbiol. Lett. **140,** 233 (1996).
19. L. E. Rodriguez-Saona, F. M. Khambaty, F. S. Fry, and E. M. Calvey, J Agric. Food Chem. **49,** 574 (2001).
20. H. M. Al-Qadiri, M. Lin, A. O. Cavinato, and B. A. Rasco, Int. J. Food Microbiol. **111,** 73 (2006).
21. V. Enikhimovitch, V. Pavlov, M. Talyshinsky, Y. Souprun, and M. Huleihel, J. Pharm. Biomed. Anal. **37,** 1105 (2005).
22. N. S. Foster, S. E. Thompson, N. B. Valentine, J. E. Amonette, and T. J. Johnson, Appl. Environ. Microbiol. **58,** 203 (2004).
23. J. Irudayaraj, H. Yang, and S. Sakhamuri, J. Mol. Struct. **606,** 181 (2002).
24. C. A. Rebuffo-Scheer, C. Kirschner, M. Staemmler, and D. Naumann, J. Microbiol. Methods **68,** 282 (2007).
25. C. A. Rebuffo-Scheer, J. Schmitt, and S. Scherer, Appl. Environ. Microbiol. **73,** 1036 (2007).
26. A. Oust, B. Moen, H. Martens, K. Rudi, T. Naes, C. Kirschner, and A. Kohler, J. Microbiol. Methods **65,** 573 (2006).
27. C. Rubio, C. Ott, C. Amiel, I. Dupont-Moral, J. Travert, and L. Mariey, J. Microbiol. Methods **64,** 287 (2006).
28. T. Udelhoven, D. Naumann, and J. Schmitt, Appl. Spectrosc. **54,** 1471 (2000).
29. C. L. Winder, E. Cam, R. Goodacre, and R. Seviour, J. AppI. Microbiol. **96,** 328 (2004).
30. H. Oberreuter, H. Seiler, and S. Scherer, Int J. Syst Evol. Microbiol. **52,** 91 (2002).
31. S. H. Beattie, C. Holt, D. Hirst, and A. G. Williams, FEMS Microbiol. Lett. **164,** 201 (1998).
32. M. C. Curk, F. Peladan, and J. C. Hubert, FEMS Microbiol. Lett. **123,** 241 (1994).
33. R. Goodacre, B. Shann, R. J. Gilbert, E. M. Timmins, A. C. McGovern, B. K. Alsberg, D. B. Kell, and N. A. Logan, Anal. Chem. **72**, 119 (2000).
34. M. Kümmerle, S. Scherer, and H. Seiler, Appl. Environ. Microbiol. **64,** 2207 (1998).
35. D. J. M. Mouwen, M. J. B. M. Weijtens, R. Capita, C. Alonso-Calleja, and M. Prieto, Appl. Environ. Microbiol. **71,** 71 (2005).
36. D. L. Perkins, C. R. Lovell, B. V. Bronk, B. Setlow, P. Setlow, and M. L. Myrick, Appl. Spectrosc. **58,** 749 (2004).
37. D. L. Perkins, C. R. Lovell, B. V. Bronk, B. Setlow, P. Setlow, and M. L. Myrick, Appl. Spectrosc. **59,** 893 (2005).
38. M. V. Schiza, D. L. Perkins, R. J. Priore, B. Setlow, P. Setlow, B. V. Brook, D. M. Wong, and M. L. Myrick, Appl. Spectrosc. **59,** 1068 (2005).
39. N. A. Baldauf, L. A. Rodriguez-Romo, A. E. Youssef, and L. E. Rodriguez-Saona, Appl. Spectrosc. **60,** 592 (2006).
40. H. Li and C. P. Tripp, Appl. Spectrosc. **62**, 62 (2008).
41. M. Theriault, E. Puckrin, and J. O. Jensen, Appl. Opt. **42,** 6696 (2003).
42. S. E. Thompson, N. S. Foster, T. J. Johnson, N. B. Valentine, and J. E. Amonette, Appl. Spectrosc. **57**, 893 (2003).
43. T. J. Johnson, Y. F. Su, N. B. Valentine, H. W. Kreuzer-Martin, K. L. Wahl, S. D. Williams, B. H. Clowers, and D. S. Wunschel, Appl Spectrosc. **63,** 899 (2009).
44. A. C. Samuels, A. P. Snyder, D. K. Emge, D. St. Amant J. Minter, M. Campbell, and A. Tripathi, Appl. Spectrosc. **63,** 14 (2009).
45. L. Mariey J. P. Signolle, C. Amiel, and J. Travert, Vib. Spectrosc. **26,** 151(2001).
46. M. A. M. Gomez, M. A. Brains Perez, F. J. M. Gil, A. D. Diez. J. F. M. Rodriguez, A. O. Domingo, and A. D. Tones, J. Microbiol. Methods **55,** 121 (2003).
47. J. Kirkwood, S. F. Al-Khaldi, M. M. Mossoba, J. Sedman, and A. A. Ismail, Appl. Spectrosc. **58,** 1364 (2004).
48. H. M. Al-Qadiri, M. Lin, A. G. Cavinato, and B. A. Rasco, Intl. J. Food Microbiol. **111,** 73 (2006).
49. N. S. Foster, S. E. Thompson, N. B. Valentine, J. E. Amonette, and T. J. Johnson, Appl. Spectrosc. **58,** 203 (2004).
50. U. Neugebauer, U. Schmid, K. Baumann, W. Ziebuhr, S. Kozitskaya, V. Deckert, M. Schmitt, and J. Popp, Chem. Phys. Chem. **8,** 124 (2007).
51. T. Kohonen, *Self-Organizing Maps* (Springer, New York, 2001), 3rd ed.
52. J. G. Dy and C. E. Brodley, J. Machine Learning Res. **5,** 845 (2004).
53. B. K. Lavine, C. E. Davidson, and D. J. Westover, J. Chem. Inf. Comput. Sci. **44,** 1056 (2004).
54. I. Kuzmanovski, M. Trpkovska, and B. Šoptrajanov J. Mol. Struct. **744,** 833 (2005).
55. H. Yin and N. M. Allinson, Neural Comput. **7,** 1178 (1995).
56. L. Hamel and C. W. Brown, "Improved Interpretability of the Unified Distance Matrix with Connected Components," *Proceedings of the 2011 International Conference on Data Mining* (Mesa, AZ, April 28–30 2011), to appear.
57. M. Cottrell, E. De Bodt, and M. Verleysen, "A statistical tool to assess the reliability of self-organizing maps," in *Advances in Self-Organising Maps* (Springer-Verlag, Amsterdam, 2001), pp. 7–14.
58. A. Bharathi and A. M. Natarajan J. Theo. Appl. Info Tech. **9,** 162 (2009).
59. K. J. Johnson and R. E. Synovec, Chemom. Intell. Lab. Syst. **60,** 225 (2002).
60. P. B. Harrington, N. E. Vieira, J. Espinoza, J. K. Nien, R. Romero, and A. L. Yergey, Anal. Chim. Acta **544,** 118 (2005).
61. R. C. Pinto, V. Bosc, H. Nocairi, A. S. Barros, and D. N. Rutledge, Anal. Chim. Acta **629,** 47 (2008).
62. A. K. Kniggendorf, T. W. Gaul, and M. Meinhardt-Wollweber, Appl. Spectrosc. **65,** 165 (2011).
63. X.-Y. Wang, J. M. Garibaldi, B. Bird, and M. W. George, "Novel Development in Fuzzy Clustering for Classification of Cancerous Cells using FTIR spectroscopy," in *Advances in Fuzzy Clustering and Its Applications*, W. Pedrycz, Ed. (John Wiley and Sons, New Jersey, 2007), pp. 405.