HHS Public Access

Author manuscript

J Pharm Sci. Author manuscript; available in PMC 2019 June 21.

Published in final edited form as:

J Pharm Sci. 2017 November ; 106(11): 3270–3279. doi:10.1016/j.xphs.2017.07.013.

Comparative Characterization of Crofelemer Samples Using Data Mining and Machine Learning Approaches With Analytical Stability Data Sets

Maulik K. Nariya[1], **Jae Hyun Kim**[2], **Jian Xiong**[2,3], **Peter A. Kleindl**[2], **Asha Hewarathna**[2], **Adam C. Fisher**[4], **Sangeeta B. Joshi**[2,3], **Christian Schöneich**[2], **M. Laird Forrest**[2], **C. Russell Middaugh**[2,3], **David B. Volkin**[2,3], and **Eric J. Deeds**[5,6,7,*]

[1]Department of Physics and Astronomy, University of Kansas, Lawrence, Kansas 66045

[2]Department of Pharmaceutical Chemistry, University of Kansas, Lawrence, Kansas 66045

[3]Macromolecule and Vaccine Stabilization Center, University of Kansas, Lawrence, Kansas 66045

[4]Center for Drug Evaluation and Research, Office of Pharmaceutical Quality, U.S. Food and Drug Administration, Silver Spring, Maryland 20993

[5]Department of Molecular Biosciences, University of Kansas, Lawrence, Kansas 66045

[6]Center for Computational Biology, University of Kansas, Lawrence, Kansas 66045

[7]Santa Fe Institute, Santa Fe, New Mexico 87501

Abstract

There is growing interest in generating physicochemical and biological analytical data sets to compare complex mixture drugs, for example, products from different manufacturers. In this work, we compare various crofelemer samples prepared from a single lot by filtration with varying molecular weight cutoffs combined with incubation for different times at different temperatures. The 2 preceding articles describe experimental data sets generated from analytical characterization of fractionated and degraded crofelemer samples. In this work, we use data mining techniques such as principal component analysis and mutual information scores to help visualize the data and determine discriminatory regions within these large data sets. The mutual information score identifies chemical signatures that differentiate crofelemer samples. These signatures, in many cases, would likely be missed by traditional data analysis tools. We also found that supervised learning classifiers robustly discriminate samples with around 99% classification accuracy, indicating that mathematical models of these physicochemical data sets are capable of identifying even subtle differences in crofelemer samples. Data mining and machine learning techniques can thus identify fingerprint-type attributes of complex mixture drugs that may be used for comparative characterization of products.

*Correspondence to: Eric J. Deeds (Telephone: +1-785-864-1057; Fax: +1-785-864-5558)., deeds@ku.edu (E.J. Deeds).

Keywords

crofelemer; comparative characterization; data mining; supervised learning

Introduction

Drugs derived from natural biological sources can be highly heterogeneous and structurally complex, if a single component is not highly purified in the manufacturing process. Naturally derived complex mixture drugs, especially from botanical sources, can exhibit batch-to-batch variation and be sensitive to changes in the manufacturing process.[1] For this reason, analytical characterization is an important element of product development.[2–4] Comparative characterization is necessary to support any postapproval changes in the manufacturing process, including potential changes in raw material sources or process parameters.

An early step in process development generally involves identifying critical quality attributes (CQAs), which are a set of physical, chemical, and biological properties associated with the drug in question.[5] Based on the nature of a CQA and the state of technology, either single or multiple analytical techniques measure the CQA, for example, during manufacturing and in stability studies. A combination of absorption spectra, chromatography, and mass spectrometry was used to monitor the CQAs of crofelemer, as described in the preceding 2 articles in this 3-article series.[6,7] CQAs are often used for comparisons between drug product batches. However, such comparisons often lack the capacity to deal with data sets of extreme size, or with the combination of orthogonal techniques to monitor a single CQA, which may be necessary to compare highly heterogeneous products. Although difficult using traditional methods of data analysis, machine learning and data mining offer a feasible approach to evaluate large and combined data sets.

In recent work on the use of IgG1-Fc glycoforms as a model system for biosimilarity assessments,[8–11] Kim et al. implemented a machine learning approach to identify potential differences between IgG1-Fc glycoform samples. In particular, the physical stability profile of eight different samples of four well-defined IgG1-Fc glycoforms in two different formulations, with three replicates each, was generated by subjecting them to different temperature and pH conditions. They found that combining physicochemical data sets from multiple experimental sources allowed them to robustly discriminate various samples using various machine learning approaches.

In this work, we selected a botanical complex mixture drug, crofelemer (Fulyzaq®, a purified oligomeric proanthrocyanidin), as a model system to generate analytical data sets and develop an integrated mathematical model for comparative characterization. Unlike the fairly well-defined IgG1-Fc glycoforms, the crofelemer biopolymer is a botanical drug substance extracted from the sap of the Croton lechleri tree. It is used for treating noninfectious diarrhea in HIV/AIDS patients undergoing antiretroviral therapies.[12] As described in the companion articles in this series of 3 articles,[6,7] we prepared different samples of crofelemer by extracting it from the drug product tablet, dissolving it in aqueous solution, and then fractionating it. Finally, each sample was incubated at 2 different

temperatures. The goal in this case was to generate crofelemer samples that were treated in slightly different ways, with the expectation that the resulting materials would exhibit subtle analytical differences. Various experiments were performed to determine the physical, chemical, and biological characteristics of each crofelemer sample. We used data mining and data visualization tools such as principal component analysis (PCA) and mutual information score to extract useful information from these data sets. We then used machine learning classification to distinguish these samples from one another. We found that by combining information-rich regions of the data sets from the various analytical experiments, we were able to distinguish the crofelemer samples from one another with very high accuracy.

Materials and Methods

Sample Preparation

As described in detail in the 2 companion articles,[6,7] for the purpose of this study we dissolved surface-scraped Fulyzaq tablets in water and obtained purified crofelemer by centrifuging the mixture. We then fractionated the filtrate using molecular weight cutoff centrifugal filters. The 10-kDa molecular weight cutoff separates the mixture into 2 parts: the 10-kDa bottom fraction, which contains molecules less than 10 kDa, and the 10-kDa top fraction, which contains molecules greater than 10 kDa. Similarly, we obtained 3-kDa top and 3-kDa bottom fractions. We also used a set of unfractionated samples for this study. The concentrations of these 5 samples were measured using a HP-8453 UV-Vis photodiode array spectrometer (Agilent Technologies, Santa Clara, CA) equipped with deuterium (D2) and tungsten (W) lamps in 1-cm path length cuvettes. The extinction coefficient (ε) used was 7.6 mL/(mg$cm) at 280 nm; details on how this extinction coefficient was determined may be found in our companion article.[6] These 5 aqueous fractions were then subjected to a stability study in which they were maintained at 2 different temperatures, 25°030C and 40°C, for 0 and 2 days and 1 week and 1 month, producing 35 distinct samples. These samples were then analyzed by the methods described below.

Biological Characterization Data Sets

The assay used by the manufacturer to measure biological activity and potency of crofelemer is proprietary information, although a redacted FDA document indicates that it is a cell-based assay.[13,14] Previous publications have used single-cell patch-clamp assays to monitor crofelemer activity,[15] but these assays are highly labor-intensive and difficult to apply to the number of different samples studied here (see below). We thus used the T84 chloride channel assay for biological characterization of the crofelemer stability samples,[16] because it can be applied in a high-throughput 96-well plate format, as described in detail in the first of the 3 articles[6] in this series. T84 cells are colon carcinoma cells which express both calcium-activated chloride channels (CaCCs) and the cystic fibrosis transmembrane conductance regulator (CFTR), another Cl^- ion channel. Crofelemer acts to selectively inhibit CaCC and CFTR chloride channels on the apical side of the intestinal wall.[15,17] A Cl^- ion quenched fluorescent dye N-(ethoxycarbonylmethyl)-6-methoxyquinolinium bromide (MQAE) was used to measure intracellular chloride of the T84 monolayers in the presence of crofelemer and its degraded forms. Inhibition of CaCC and CFTR channels in the presence of crofelemer was challenged using ionomycin and forskolin, respectively. The samples for this

assay were organized slightly differently than described in the previous section. Specifically, the bottom fractions were excluded in this study due to limited availability of the lot of Fulyzaq used in the forced degradation study. We also did not use the day 2 stability time point in this assay due to limited availability of the lot of crofelemer active pharmaceutical ingredient we could obtain. On an individual plate, we ran the T84 assay with each crofelemer sample alone and in combination with ionomycin, an ionophore that facilitates transfer of Ca^{++} ion into the cell, which results in activation of CaCC, and forskolin, which activates adenylyl cyclase and results in higher intracellular cyclic AMP levels. Cyclic AMP increases protein kinase A activity which results in activation of the CFTR, another Cl^{-} ion channel. In addition to these, we also ran the assay with no treatment as a negative control and with the flavonoid quercetin as a positive control. As a result, there are 75 instead of 35 samples for this assay. To minimize experimental uncertainties associated with the plate readings, we ran replicates of these samples on 3 different plates and ran 8 replicates of the samples on each of these plates to obtain a greater statistical significance. The assay involves measuring the Cl^{-} ion flux as a function of time for different samples treated with the previously mentioned examples. This gives us a total of 1800 trajectories, and for each one of those trajectories, we have measurements at 60 different time points, resulting in a total of 108,000 determinations.

Physicochemical Characterization Data Sets

To characterize the physicochemical properties of the crofelemer stability samples, a wide range of techniques including UV-visible absorption spectroscopy (UV-Vis), Fourier transform infrared spectroscopy (FTIR), circular dichroism (CD), and HPLC techniques such as size-exclusion chromatography (SEC) and hydrophilic interaction chromatography (HILIC) were used as described in detail in the 2 companion articles.[6,7] The raw data obtained from these experiments were normalized using the concentration for the 5 stability samples: 3-kDa bottom, 3-kDa top, 10-kDa bottom, 10-kDa top, and unfractionated. There are 4 replicates of each of these stability samples in all experiments. The data for UV-Vis spectroscopy report the absorbance value for different wavelengths ranging from 190 to 1100 nm. The organization of the FTIR and CD data is very similar to that of UV-Vis, the only difference is the different ranges of wavelengths; FTIR measures absorbance between 800 and 4000 cm^{-1} and CD measures the difference in absorbance between the right and the left circularly polarized light of wavelengths between 200 and 350 nm. SEC and HILIC separate the sample over a size-based or hydro-phobic interactionebased column, respectively, and record UV spectra for each retention time. For CD, HILIC, and SEC, the experiment was run with buffer (without any sample treatment), and we subtracted these background values from the raw data before normalizing it by concentration. Some of these techniques had artifactual data in some ranges of wavelengths. Thus, we did not include the data from these regions in our analysis. We included the wavelengths between 240 and 600 nm for UV, wavenumber between 1100 and 1700 cm^{-1} for FTIR, and the retention times between 10 and 40 min for all wavelengths for HILIC.

Principal Component Analysis

It is often helpful to visualize data in 2-D space for a more intuitive understanding of the data set. PCA is a commonly used dimensionality reduction technique that we applied to the

data. PCA transforms the data in such a way that new basis vectors are arranged in the order of the amount of the variance that they capture. In other words, the first principal component is calculated such that the projection of the data on to that component has the maximum variance. To perform the PCA of a given data set, the data are organized in the form of a matrix such that the rows represented the different samples and the columns represented different features of these samples. It is a common practice in data processing to rescale the data so that each column has zero mean and unit variance. This method is known as “feature scaling” or “standardization” and it helps to combine features from different experimental techniques that may have a broad range of values compared to others. We used the function scale() from sklearn.preprocessing for standardization and PCA().fit().transform() from sklearn.decomposition in Python to calculate the first 3 principal components for the data.

Mutual Information Score

Mutual information is a data mining technique that we used to identify regions of the data set that are rich in information content. The mutual information between 2 discrete random variables is defined as follows[18]:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

where X and Y are sets of possible x and y bins, $p(x,y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability of distribution functions of X and Y, respectively. For our data, the y-bins are the 35 categorical variables representing the different samples and the x-bins estimate the probability of a particular range of feature values.[7] $I(X, Y)$ is a measure of statistical correlation between the random variables. In other words, higher values of I mean that knowing the value of a particular measurement (e.g., absorbance at 350 nm in the UV-Vis experiment) significantly reduces the uncertainty in which sample (e.g., unfractionated 7 days at 40°C vs. 10-kDa top 2 days at 25°C) was measured. We used the function histogram2d() from numpy to create the 2D histograms with 6 x-bins and 35 y-bins, and we used mutual_info_score() from sklearn.metrics in Python to calculate the mutual information score.

Classification

For this part of the analysis, we made use of some of the most commonly supervised learning algorithms to classify the data, namely k-nearest neighbors (kNN), support vector machines (SVMs), decision trees (DTs), AdaBoost, random forest (RF), naive Bayes’, and linear discriminant analysis (LDA). Some of these classifiers are nonparametric whereas others have a number of parameters that must be tuned for them to be suitable for our problem. kNN classifiers work by calculating the Euclidean distance between the data points and classifies a given point based on a majority vote of its kNN.[19] SVMs construct a hyperplane that maximizes the size of the boundary between classes in a high-dimensional space; this hyperplane is then used for classification.[20] We tried 2 different kernels for our SVMs: the linear and the radial basis functions. DTs map the observations about a data point to conclusions about its target value. AdaBoost training is a special kind of decision tree which selects those features which are known to improve the classification efficiency, in

terms of both speed and accuracy.[21] RFs fall under ensemble learning methods of classification and construct a multitude of decision trees at training time and output the class that is the mode of the prediction across all of these trees.[22] The naive Bayes' classifier is a probabilistic classifier based on applying Bayes' theorem with the assumption that the features are independent of each other.[23] LDA finds a linear combination of features that separate the classes from one another.[24,25]

Cross-Validation and Classification Accuracy

It is common to validate classification analysis by using cross-validation to assess how the classifier will generalize to an independent data set and help reduce the effect of problems such as overfitting. Cross-validation involves partitioning a sample of the data into training and test sets, performing the analysis on the training set, making predictions about the classes by using this model on the test set, and calculating the accuracy by looking at the frequency of correctly made predictions.[26]

There are several techniques that can be used to achieve this goal which are broadly categorized into 2 types: exhaustive and nonexhaustive cross-validation. Exhaustive cross-validation methods make all possible combinations of the training and test data subsets. Leave-p-out is the most general form of an exhaustive cross-validation technique, which would evaluate a learning model for C_p^n combinations, where n is the total number of samples of training and test data sets. It is easy to see that even for moderately large n, leave-p-out becomes too excessive to calculate. Non-exhaustive cross-validation methods do not compute all the possible combinations of training and test data sets and are an approximation to leave-p-out cross-validation. One of the common techniques is k-fold cross-validation, where the original data set is randomly partitioned into k equal-sized sub-data sets; one of the k subsets is treated as a test set whereas the others are used as the training set. This process is then repeated using the other k-1 subsets for testing and the average accuracy is calculated. A variation of k-fold cross-validation, known as Monte Carlo cross-validation, randomly splits the data into test and training sets as well. The difference in this case is that the proportion of test to train sets is not dependent on the number of iterations. We used this technique for cross-validation of our data; specifically, we split the data into 80% training set and 20% test set and repeated this process 100 times to calculate the accuracy for a given classifier.

We optimized the parametric classifiers by calculating the classification accuracy for a range of parameters. The number of neighbors k for the kNN classifier was varied between 1 and 5. We varied the parameters C and gamma for both the linear and radial basis function SVMs. For DTs and RF, we varied the maximum number of features under consideration when looking for the best split. In addition to that, we also varied the maximum depth and minimum sample split for the tree for DTs and number of estimators for RF and AdaBoost classifiers. We performed Monte Carlo cross-validation described above for each possible combination of the parameters (depending on the classifier) and chose the parameters that resulted in highest average classification accuracy. Table 1 summarizes the parameter values we used to optimize our classifiers.

Results

Analysis of Bioassay Data

It is evident from Figure 1a bioassay data for the various crofelemer stability samples that there is a considerable overlap between the trajectories from different samples. The sheer volume of the data in this case (1800 trajectories with a total of 108,000 measurements) makes it difficult to quantitatively evaluate this overlap from the raw data. We thus used PCA to reduce dimensionality of the data. To perform PCA on the bioassay data, we organized it in the form of a matrix such that the rows represented the different replicates of the sample treatments and the columns represented the fluorescence values at a given time. Figure 1b shows the first 2 principal components. Note that the first principal component captures over 99% of the variance, indicating the bioassay data maps naturally to a 1-dimensional space. We have retained the second principal component in Figure 1b for visual clarity.

We can see that quercetin, which is the positive control, is very well separated from the rest of the sample data, but the remaining treatments, including blank, which is the negative control, have significant overlap between samples. The only exceptions are the 10-kDa 30-day sample and the unfractionated 7-day samples, which show some clustering. We found that the samples that did cluster (e.g., quercetin control) did so because they had, on average, lower (or higher) fluorescence intensities than the rest of crofelemer stability samples. This suggested that normalizing the data might permit greater separation among bioassay data from the samples. To do this, we tried the following normalization schemes for each plate:

$$I'(t) = \frac{I(t)}{I_{\max}}$$

and

$$I''(t) = \frac{I(t) - I_{\min}}{I_{\max}}$$

where $I(t)$ is the raw intensity for a given replicate at time t, and $I_{\min}$ and $I_{\max}$ are the minimum and maximum intensities observed on a plate, respectively. In all the cases, we observed that the first principal component still captured more than 99% of the variance in the data, and neither normalization scheme generated any greater separation between the samples than we observed with the raw data (see Fig. 1b).

Because PCA was unable to capture any distinguishing features of the bioassay data with the crofelemer fractionated and stability samples, we tried a different dimensionality reduction method. Figure 1c shows a representative trajectory, which seems to follow an exponential relaxation. We thus tried a 3-parameter fit to the trajectories of these samples using the function:

$$I(t) = I_0 + \Delta I\left(1 - e^{-kt}\right),$$

where I_0 denotes the initial intensity, ΔI represents the difference between initial and final intensities, and k is the rate parameter of the exponential relaxation. We found that most of the trajectories from the bioassay data fit this function well. Figures 1d–1f show the scatter plots between the fit parameters. It is evident from these scatter plots that I_0 shows differences between quercetin control, 10-kDa 30-day, and unfractionated 7-day crofelemer samples, which is similar to the findings from the PCA. There are, however, no significant differences in k or ΔI, indicating that there are no large differences among treatments (even between quercetin and blank) in terms of influence of the treatment on Cl^- efflux rate. From our analysis, we conclude that the T84 Cl channel assay lacks the requisite precision to distinguish between the majority of the different crofelemer treatments (various fractionated stability samples) under these conditions.

Analysis of UV-Vis, FTIR, CD Data

As described in Materials and Methods, each of the crofelemer stability samples was subjected to UV-Vis absorption, FTIR, and CD analyses. After background correction and normalization (as appropriate, see Materials and Methods), we found that, unlike the bioassay data, all of these biophysical techniques exhibited significant and reproducible variation between samples of different types (see Fig. 2). To quantify and visualize this result, we used mutual information scores: specifically, we independently calculated the mutual information between the signal at each wavelength in each experiment and the sample type (see Materials and Methods). We saw that each technique had regions with relatively high mutual information score (1.5–2 bits). It is thus clear that UV-Vis absorption, FTIR, and CD analyses can discriminate the crofelemer stability samples from each other, at least to some extent, which indicates that they should represent useful inputs for classification analyses (see below).

Analysis of HILIC and SEC Data

The HPLC techniques (i.e., SEC and HILIC) have much larger data sets, because they obtain a full UV absorption spectrum (as shown in Fig. 2a) for each retention time across the chromatogram. We thus have a total of 1,395,744 measurements for SEC and 2,326,240 measurements for HILIC for each crofelemer stability sample. It is impractical in this case to use these entire data sets as features for classification. We thus employed the mutual information score approach not only to examine the general discriminatory power of these data but also to find useful subsets of the chromatography data for further analysis for comparative characterization of the various crofelemer samples.

The heat map of mutual information scores for SEC is shown in Figure 3a. Traditionally, it is common to choose one wavelength (e.g., 280 nm) and plot the absorbance at that wavelength across all retention times. As one can observe in Figure 3b, however, the area of highest mutual information scores is spread across a wide range of wavelengths for a narrow set of retention times. We quantified this observation by taking the average of all mutual

information scores over all wavelengths for a given retention time or over all retention times for a given wavelength (see Fig. 3b). Note that the average mutual information is highest for retention times between 10 and 12 min, averaged over all wavelengths. Figure 3c shows the absorbance values across all wavelengths for the retention time with the highest average mutual information score (12.1 min), and we can see that there are very significant differences between the samples.

From the mutual information score heat map of the HILIC data obtained from the crofelemer stability samples (Fig. 3d), we observed that there was more information content in the conventional "slice" of the data set, that is, a range of wavelengths (between 260 and 300 nm) across all retention times. Just as in the SEC case, we averaged the mutual information scores over all wavelengths and over all times for HILIC data sets (see Fig. 3e). For HILIC, the data subset consisting of absorbance values for the wavelength near 282 nm had the maximum average mutual information score across all retention times. The absorbance values for these retention times at approximately 282 nm show that there are indeed significant differences between the samples (see Fig. 3f). Despite the rather traditional wavelength identified for HILIC, these results highlight the fact that mutual information scores are a useful tool for data mining, that is, finding informative "slices" of these extremely high dimensionality data sets, some of which might easily be missed in a more traditional experimental analysis.

Data Visualization of the Physical and Chemical Analysis

Our mutual information score analysis suggested that the combination of the following data sets could be useful in classifying the crofelemer stability samples: UV-Vis absorption (240–600 nm), CD (200–350 nm), FTIR (1100–1700 cm^{-1}; see Figs. 2a–2c), SEC for all wavelengths at retention times near 12.1 min (see Figs. 3a–3c), and HILIC at all retention times at wavelength near 282 nm (see Figs. 3d–3f). This resulted in a total of 4441 features per sample. To visualize the organization of the data in this high-dimensional space, we performed PCA (see Fig. 4). We found that there is generally very good separation among classes. Not only is the average separation between crofelemer stability samples better than the bioassay data (see Fig. 1b), but the variance is also spread out more evenly across the principal components (the first 2 components collectively capture less than 80%). Overall, this indicates that the physicochemical data sets have much better discriminatory power than the bioassay data.

Classification Analysis

From the mutual information calculation, we saw that all the techniques on their own have individual data points with a relatively high mutual information score (~2 bits); while this is encouraging, it is clear that no single feature can completely discriminate among these sample types. We thus need to combine these features to classify the samples; this represents a classic problem of "supervised" machine learning. We thus applied a set of standard classifiers not only using all of the techniques individually but also for all possible pairs, combination of 3, combination of 4, and finally all 5 analytical techniques combined (see Materials and Methods). We used the Monte Carlo cross-validation technique, splitting the data into 80% training and 20% testing sets, and averaged the classification accuracy of our

classifiers across 100 such instants to obtain the accuracies reported in Table 2 (see Materials and Methods).

Table 2 summarizes the results of the classification accuracy for different classifiers for all possible combinations of the techniques in terms of being able to discriminate each of the crofelemer stability samples from each other. For a single technique, the UV absorption data with the LDA classifier performs best, with over 97% classification accuracy. The UV absorption data also performs well with kNN, SVMs (both linear and radial basis function kernel), and RF, with a classification accuracy close to 95% in each case. We found that data from the other techniques did not perform nearly as well when used in isolation. HILIC + UV and HILIC + CD + UV are the combinations with best classification accuracy (over 98%).Interestingly, the combination of HILIC + CD + UV data has a classification accuracy higher than the ones in all combinations of 4 and 5 techniques, demonstrating that it may not always be better to have more features for classification. Top performers for any number of combinations of techniques include the UV absorption data and the LDA classifier. Also note that while these combinations improve performance, they are not all that much better than the UV absorption data alone with the LDA classifier (ca. 97% vs. ca. 98% each). The top-performing combinations that do not include UV are SEC + FTIR and SEC + HILIC (over 96% classification accuracy).

The LDA classifier consistently outperformed the others on nearly every data combination (Table 2). Nonetheless, a variety of methods produce acceptable accuracy (>95%) using various combinations of the techniques. Overall, these finding indicate that the physicochemical data sets contain enough information to robustly classify each of the different crofelemer stability samples from each other. We should note that Kim et al.[8] found 100% classification accuracy for an analytical data set focused on IgG1-Fc stability. Given the smaller sample sizes available in that previous study, the authors employed leave-1-out and leave-8-out cross-validation, resulting in smaller test sets than employed here. We found multiple examples of 100% classification accuracy in our analysis using leave-1-out cross-validation (data not shown), indicating that the slightly lower accuracies we obtained were likely due to the cross-validation scheme we employed, rather than the fact that the IgG1-Fc samples were more structurally well defined.

Discussion

Crofelemer is 1 of 2 FDA-approved botanical drug products. Botanical drugs harbor more heterogeneity than typical small-molecule drugs owing in part to the biological variability of the naturally derived raw material. For botanical drug products, analytical characterization is important for product development. Comparative characterization supports any postapproval changes to the manufacturing process. In this work, we used the botanical drug crofelemer as a model system for comparative characterization of a therapeutic complex mixture drug. Unlike protein therapeutics, crofelemer is a polymeric natural mixture of compounds both heterogeneous in size and less well defined chemically and structurally than a protein. We used 5 different fractions of crofelemer and subjected each of them to different temperatures for different lengths of time, as described in the 2 companion articles in this series of 3 articles,[6,7] as a model system to mimic different lots of crofelemer. We thus obtained 35

distinct samples, which were then subjected to biological, chemical, and physical characterization and analysis. Our goal was to capture potentially subtle differences between these samples and assess to what extent the crofelemer stability samples can be distinguished from each other. Modern data mining and machine learning approaches offer a natural way to characterize these differences within analytical data sets. Specifically, we used data mining and visualization tools such as mutual information score and PCA to better understand such data. We found interesting differences among samples and selected features for classification analysis. We also used a standard set of supervised learning classifiers to classify crofelemer samples using the data that were available from different physical and chemical assays.

Our analysis revealed that the bioassay data obtained from a T84 chloride channel biological assay was not significant. In particular, PCA showed few significant differences between the various crofelemer treatments (see Fig. 1b). The fact that the first principal component itself captured over 99% of the variance in the data implies that the data are primarily 1-dimensional. Looking at the trajectories of the Cl^- ion efflux, we saw that the main differences between them were probably in their initial intensity. This motivated us to fit the trajectories using an exponential relaxation function. The fits to these trajectories showed no significant differences in k, which implies that different treatments had few detectable effects on the Cl^- ion efflux rate (see Figs. 1d–1f). These results indicate that the physicochemical differences between the crofelemer samples did not significantly affect the readout in this particular biological assay. Given that the T84 assay could not distinguish the activity of almost any crofelemer sample from a blank control, it is likely that the assay itself lacks the precision to capture differences in the biological activity of the degraded samples studied here. It may also be the case that the chemical differences between these samples do not translate into biologically meaningful differences. Future work, perhaps relying on more sensitive single-cell patch-clamp assays,[15] will likely be needed to determine whether degradation of crofelemer samples has an impact on biological activity.

Physical and chemical assays, particularly the HPLC techniques, generated extremely large volumes of data and it is difficult to analyze these data sets with traditional data analysis methods and identify any meaningful differences among the crofelemer stability samples. The mutual information score allowed us to identify subsets or “slices” of these large data sets that capture the most significant physical and chemical differences among the samples[7] and at the same time also helped us improve further analysis of the data. For instance, we saw that in the case of SEC data, the “traditional” analysis method of considering all retention times at a given wavelength is not the best method for detecting differences; instead looking at all of the wavelengths for the retention time of circa 12.1 min was more informative. This method allowed us to perform feature selection, which, in case of SEC and HILIC, reduced the number of features from around 3,700,000 to close to 3000 for each sample.

We obtained over 98% classification accuracy for some of the analytical data sets using the machine learning classifiers. This implies that the data sets are rich enough to robustly distinguish subtle differences between crofelemer samples by combining multiple data sets. We found that the LDA classifier outperformed the others for this data set with a

classification accuracy of over 90% for most combinations of techniques. The combinations with top accuracies all include the UV absorption data, suggesting that this information has more discriminatory capability than others. This indicates that the chemical changes that occur during crofelemer degradation are readily reflected in the vibrational modes of the covalent bonds in the material. Because many of these changes are due to chemical oxidation,[6,7] this is perhaps not surprising. Further work will be needed, however, to fully characterize the chemical changes that are generating the observed differences in the UV absorption spectra.

Overall, our results suggest that machine learning classification applied to the physical and chemical assay data sets is a promising approach for comparative characterization between complex mixture drugs such as crofelemer. Although the bioassay described here did not detect significant differences between samples, our findings indicate that physical characterization data contain sufficient information to distinguish between both subtle and overt chemical differences between samples. Such data should be capable of distinguishing between biologically active and inactive material when given a bioassay sensitive enough to make conclusive determinations regarding potency and activity. Despite the heterogeneity of crofelemer, further analysis of these types of data sets from multiple lots from different manufacturers may permit a better understanding of the structural origin of the subtle differences responsible for the differences detected. Moreover, there results indicate that data mining and machine learning analysis of various analytical data sets may be able to provide a fingerprint analysis of complex molecules and provide an integrated mathematical model for comparative characterization. Future challenges will be determining the significance and potential clinical relevance and risk of differences identified by such a model and using multiple product lots to establish the range of variation between lots from the same manufacturer.

Acknowledgments

Funding for this work was made possible by the Food and Drug Administration through grant 1U01FD005285–01. Views expressed in this publication do not necessarily reflect the views or policies of the Food and Drug Administration and the official policies of the Department of Health and Human Services nor does any mention of trade names, commercial practices, or organization imply endorsement by the U.S. Government. In addition, PAK was funded in part by R01CA173292.

References

1. Lee SL, Dou JH, Agarwal R, et al. Evolution of traditional medicines to botanical drugs. Science. 2015;347(6219):S32–S34.
2. Food US and Administration Drug. Botanical drug development: Guidance for Industry. Silver Spring, MD: Center for Drug Evaluation and Research, Food and Drug Administration; 2015.
3. Federici M, Lubiniecki A, Manikwar P, Volkin DB. Analytical lessons learned from selected therapeutic protein drug comparability studies. Biologicals. 2013;41(3):131–147. [PubMed: 23146362]
4. Lubiniecki A, Volkin DB, Federici M, et al. Comparability assessments of process and product changes made during development of two different monoclonal antibodies. Biologicals. 2011;39(1): 9–22. [PubMed: 20888784]
5. Alt N, Zhang TY, Motchnik P, et al. Determination of critical quality attributes for monoclonal antibodies using quality by design principles. Biologicals. 2016;44(5):291–305. [PubMed: 27461239]

6. Kleindl PA, Xiong J, Hewarathna A, et al. The botanical drug substance crofelemer as a model system for comparative characterization of complex mixture drugs. J Pharm Sci. 2017;106(11): 3242–3256. [PubMed: 28743606]
7. Hewarathna A, Mozziconacci O, Nariya MK, et al. Chemical stability of the botanical drug substance crofelemer: a model system comparative characterization of complex mixture drugs. J Pharm Sci. 2017;106(1):3257–3269. [PubMed: 28688843]
8. Kim JH, Joshi SB, Tolbert TJ, Middaugh CR, Volkin DB, Smalter Hall A. Bio-similarity assessments of model IgG1-Fc glycoforms using a machine learning approach. J Pharm Sci. 2016;105(2):602–612. [PubMed: 26869422]
9. More AS, Toprani VM, Okbazghi SZ, et al. Correlating the impact of well-defined oligosaccharide structures on physical stability profiles of IgG1-Fc glycoforms. J Pharm Sci. 2016;105(2):588–601. [PubMed: 26869421]
10. Mozziconacci O, Okbazghi S, More AS, Volkin DB, Tolbert T, Schoneich C. Comparative evaluation of the chemical stability of 4 well-defined immunoglobulin G1-Fc glycoforms. J Pharm Sci. 2016;105(2):575–587. [PubMed: 26869420]
11. Okbazghi SZ, More AS, White DR, et al. Production, characterization, and biological evaluation of well-defined IgG1 Fc glycoforms as a model system for biosimilarity analysis. J Pharm Sci. 2016;105(2):559–574. [PubMed: 26869419]
12. Patel TS, Crutchley RD, Tucker AM, Cottreau J, Garey KW. Crofelemer for the treatment of chronic diarrhea in patients living with HIV/AIDS. HIV AIDS (Auckl). 2013;5:153–162. [PubMed: 23888120]
13. Food US and Administration Drug. Exclusivity summary. Silver Spring, MD: FDA; 2012 Available at: https://www.accessdata.fda.gov/drugsatfda_docs/nda/2012/202292Orig1s000Admincorres.pdf. Accessed August 7, 2017.
14. Glenmark Pharmaceuticals Ltd. Manufacture and supply agreement. Available at: https://www.sec.gov/Archives/edgar/data/1585608/000110465915078779/a15-18097_1ex10d2.htm. Accessed August 7, 2017.
15. Tradtrantip L, Namkung W, Verkman AS. Crofelemer, an antisecretory antidiarrheal proanthocyanidin oligomer extracted from Croton lechleri, targets two distinct intestinal chloride channels. Mol Pharmacol. 2010;77(1):69–78. [PubMed: 19808995]
16. Food US and Administration Drug. Guidance for Industry: Q10 Pharmaceutical Quality Systems. Silver Spring, MD: Center for Drug Evaluation and Research, Food and Drug Administration and Rockville, MD; Center for Biologics Evaluation and Research, Food and Drug Administration; 2009.
17. Gabriel SE, Davenport SE, Steagall RJ, Vimal V, Carlson T, Rozhon EJ. A novel plant-derived inhibitor of cAMP-mediated fluid and chloride secretion. Am J Physiol. 1999;276(1):G58–G63. [PubMed: 9886979]
18. Cover TM, Thomas JA. Elements of Information Theory. 2nd ed Hoboken, NJ: Wiley; 2006.
19. Altman NS. An introduction to Kernel and nearest-neighbor nonparametric regression. Am Statistician. 1992;46(3):175–185.
20. Chang Y-W, Hsieh C-J, Chang K-W. Training and testing low-degree polynomial data mappings via linear SVM. J Mach Learn Res. 2010;11:1471–1490.
21. Freund Y, Schapire RE. Experiments with a new boosting algorithm Machine Learning: Proceedings of the Thirteenth International Conference. Murray Hill, NJ: AT&T Laboratories; 1996.
22. Breiman L Random forests. Mach Learn. 2001;45(1):5–32.
23. Salzberg SL. Book review: C4.5: Programs for Machine Learning by J. Quinlan Ross. Morgan Kaufmann Publishers, Inc., 1993 Available at: https://link.springer.com/content/pdf/10.1007%2FBF00993309.pdf. Accessed August 16, 2017.
24. Hubetry CJ, Olejnik S. Applied MANOVA and Discriminant Analysis. 2nd ed Hoboken, NJ: Wiley; 2006.
25. McLachlan G Discriminant Analysis and Statistical Pattern Recognition. Hoboken, NJ: Wiley; 2004.
26. Kohavi R International Joint Conference on Artificial Intelligence, 1995.

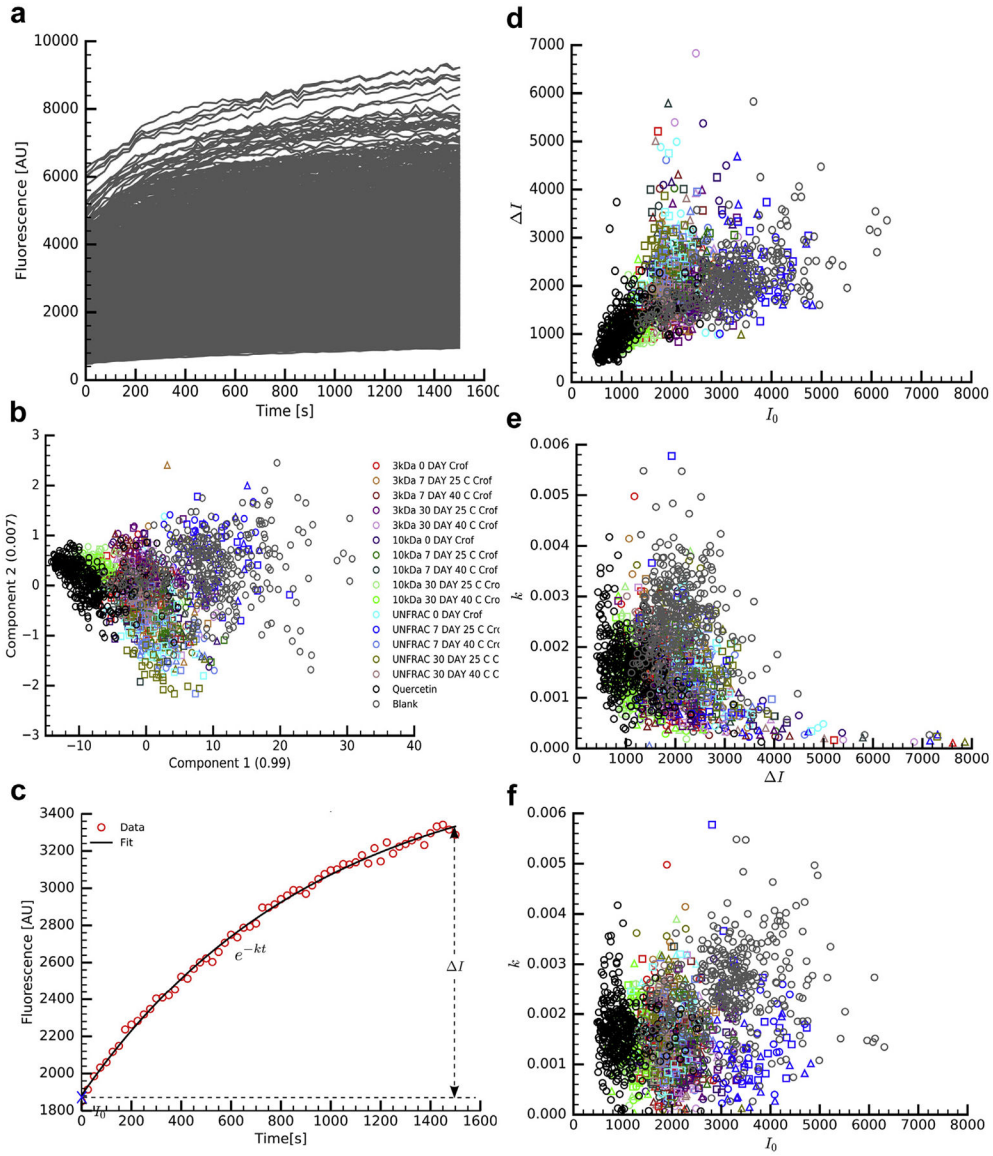

Figure 1.

Analysis of T84 chloride channel bioassay data of various crofelemer stability samples. (a) Raw data for all 1800 trajectories. (b) First and second principal components of the data. The triangles represent crofelemer treated with ionomycin whereas the squares represent crofelemer treated with forskolin. (c) An example of the fit to a typical raw trajectory. (d), (e), and (f) Scatter plots between the fitted parameters (k, ΔI, and I_0) for all 1800 trajectories.

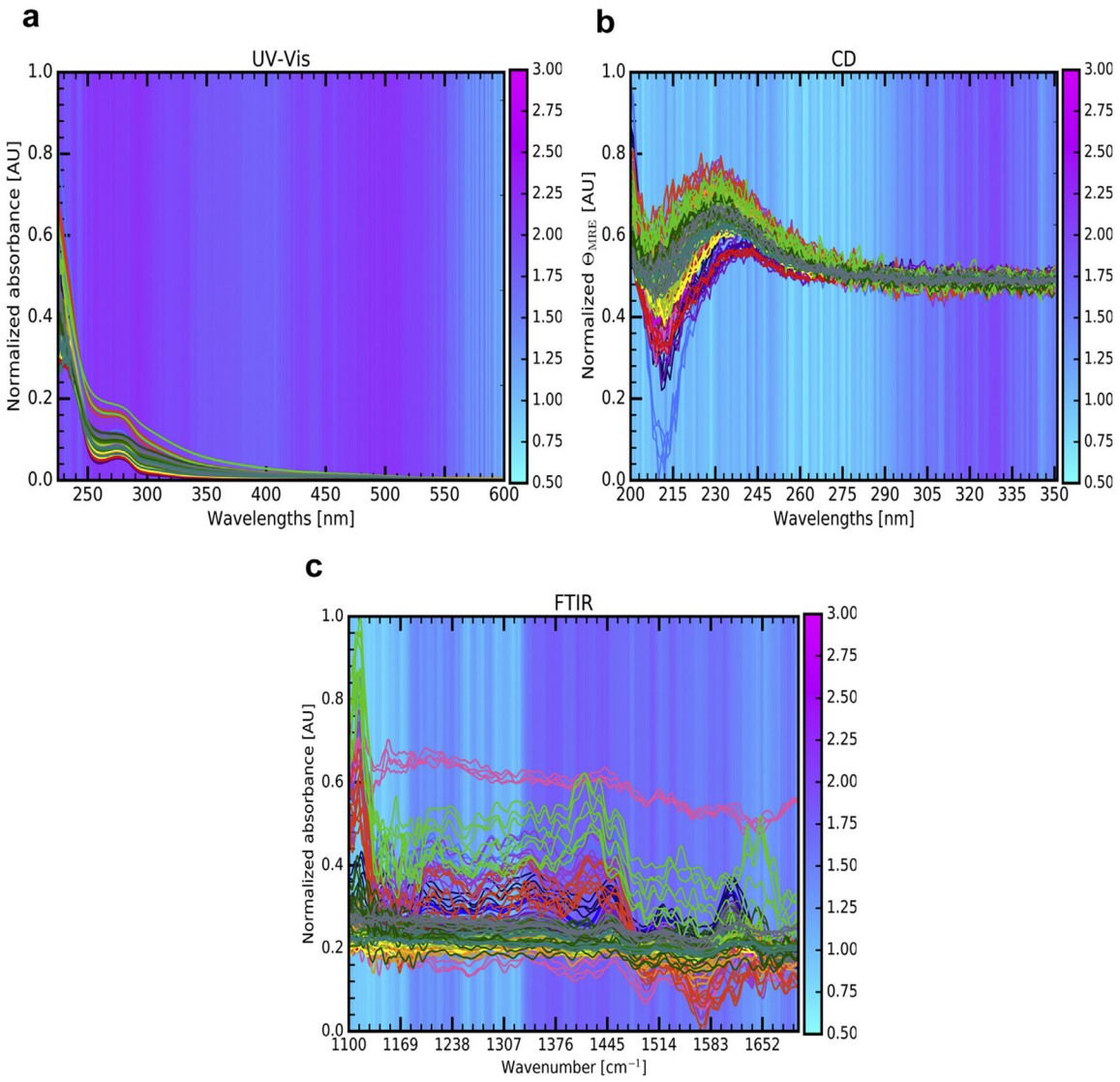

Figure 2.

Raw data and mutual information score for (a) UV-Vis absorption, (b) CD, and (c) FTIR data from various crofelemer stability samples. In each plot, the colored lines show the normalized data for the corresponding technique and the background shows the mutual information score. In each case, the raw data were divided by the concentration of the samples, and the maximum intensity in each case was normalized to 1.

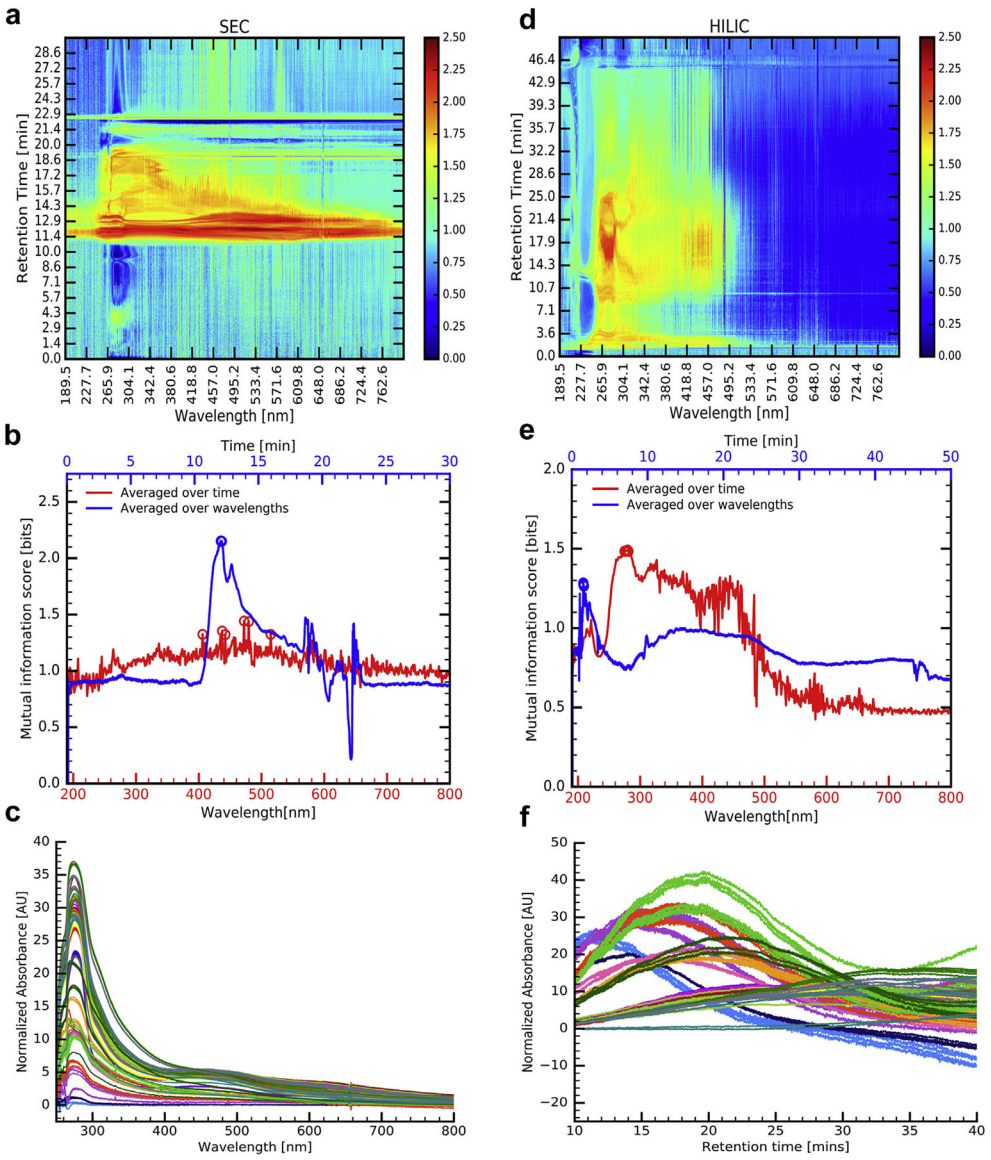

Figure 3.

Plots for mutual information score and normalized slice of data for SEC and HILIC of various crofelemer stability samples. (a) Heat map of mutual information score for SEC data. (b) The red line represents the mutual information score averaged over all retention times, and the blue line represents the mutual information score averaged over all wavelengths for SEC data. In each case, the circles represent the top 6 average mutual information score. (c) SEC data for the retention time of 12.1 min, background corrected and normalized using the dilution factors of the samples. (d) and (e) same as in (a) and (b) for HILIC. (f) HILIC data for the wavelength of 282 nm background corrected and normalized with dilution factors for the samples.

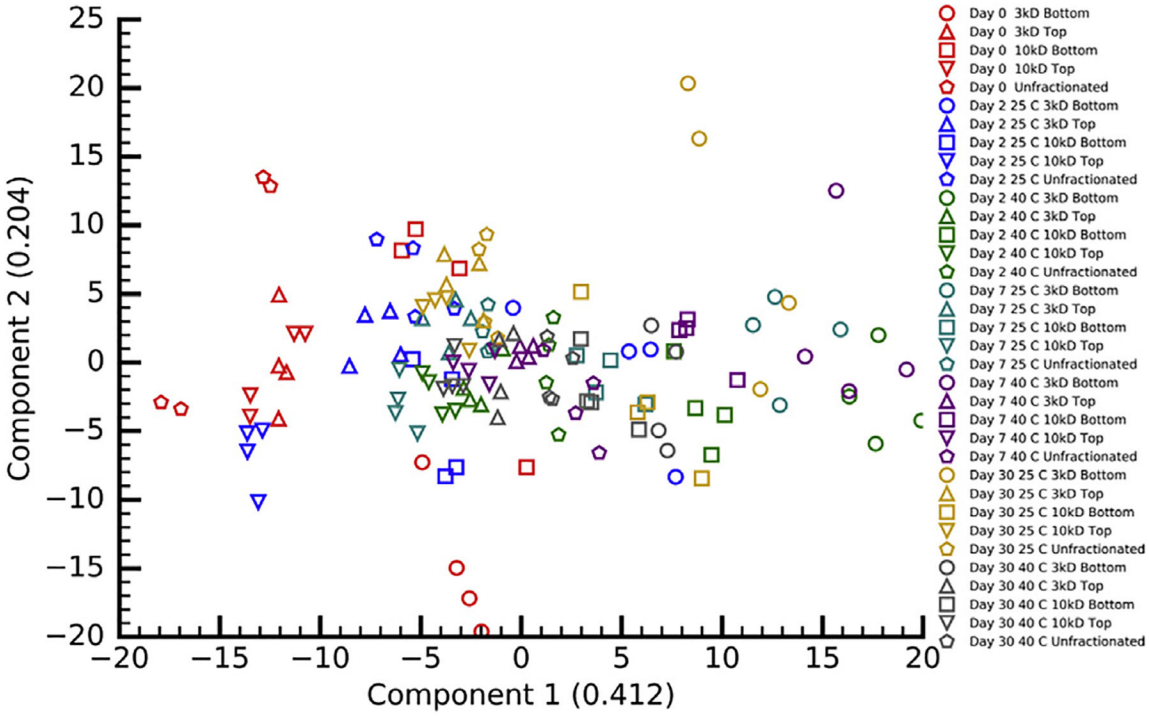

Figure 4.

PCA of combination of data from UV-Vis absorption, CD, FTIR, SEC, and HILIC analyses of various crofelemer stability samples.

Table 1

Parameters Used to Optimize Each Classifier Used in Supervised Machine Learning

Classifier	Parameter	Meaning	Tested Values
kNN	*k*	Number of neighbors	1,2, 3, 4, 5
SVM (linear)	*C*	Penalty error term	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1
	Gamma	Kernel coefficient	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1
SVM (RBF)	*C*	Penalty error term	10^{-2}, 10^{-1}, 1, 10, 100
	Gamma	Kernel coefficient	10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}
Decision tree	Max features	Max number of features	Sqrt, Log2, Max
	Max depth	Max depth of the tree	5,10, 20, 50,100, 200
	Min sample split	Min samples required to be a leaf node	1,2
Random forest	Max features	Max number of features	Sqrt, Log2, Max
	Number of estimators	Number of estimators	5,10, 20, 50,100, 200
	Min sample split	Min samples required to be a leaf node	1,2
AdaBoost	Number of estimators	Number of estimators	50, 10, 150, 200, 250, 300
Naive Bayes’	Not applicable	-	-
LDA	Not applicable	-	-

RBF, radial basis function.

Table 2

Summary of Results From Classification Analysis of Physicochemical Data From Various Crofelemer Stability Samples

Techniques	kNN	Linear SVM	RBF SVM	LDA	Naive Bayes'	Decision Tree	Random Forest	AdaBoost
HILIC	61.43	76.39	67.25	81.36	36.5	55.89	66.71	23.54
SEC	82.64	84.39	83.75	72.00	36.82	71.11	85.75	40.43
CD	35.18	42.25	39.93	45.04	12.21	29.75	48.11	9.79
FTIR	69.29	81.21	78.46	**95.32**	29.00	44.75	61.93	24.25
UV	93.82	94.18	93.11	***97.39***	68.86	86.5	93.39	58.43
HILIC + SEC	76.75	87.32	83.64	92.21	36.75	65.96	84.04	24.68
HILIC + CD	66.61	78.14	71.96	86.75	34.21	56.00	69.79	18.43
HILIC + FTIR	73.14	82.89	78.54	94.11	36.39	60.61	75.64	23.18
HILIC + UV	80.39	89.07	84.11	***98.61***	48.00	68.96	85.75	30.5
SEC + CD	71.36	76.29	74.82	82.18	33.89	66.43	86.64	21.14
SEC + FTIR	89.25	90.21	89.86	**96.46**	39.32	68.89	87.57	27.36
SEC + UV	90.21	91.79	92.29	**96.68**	54.61	80.64	93.46	38.32
CD + FTIR	67.29	77.68	76.00	94.25	30.5	48.32	70.43	19.79
CD + UV	81.50	84.54	82.57	92.36	52.89	78.18	91.50	30.57
FTIR + UV	93.96	94.46	93.57	**98.50**	54.36	75.21	90.96	35.71
HILIC + SEC + CD	76.39	86.25	82.57	92.07	37.32	64.96	86.64	20.89
HILIC + SEC + FTIR	84.43	89.89	85.18	**97.54**	40.86	66.46	86.04	24.11
HILIC + SEC + UV	85.32	90.96	87.21	**97.89**	47.57	70.11	89.71	30.71
HILIC + CD + FTIR	79.79	87.46	83.14	**95.11**	34.5	58.61	77.11	19.93
HILIC + CD + UV	80.75	88.25	84.07	***98.82***	44.82	67.79	83.93	25.89
HILIC + FTIR + UV	86.93	90.36	87.18	**98.71**	45.50	67.21	86.86	28.32
SEC + CD + FTIR	88.18	91.79	90.18	**95.32**	37.32	66.54	87.71	23.36
SEC + CD + UV	87.21	89.68	88.54	**96.71**	50.32	76.75	93.18	29.64
SEC + FTIR + UV	92.89	**95.54**	94.54	**98.00**	52.64	74.64	93.54	34.46
CD + FTIR + UV	89.71	94.14	92.29	**98.75**	49.36	73.07	89.36	25.39
HILIC + SEC + CD + FTIR	84.93	90.18	89.86	**96.04**	37.39	64.29	86.36	22.00
HILIC + SEC + CD + UV	82.82	89.75	87.04	**98.11**	46.82	69.64	89.25	25.21
HILIC + SEC + FTIR + UV	89.14	93.36	90.89	**98.25**	45.46	69.57	90.89	26.50

Techniques	kNN	Linear SVM	RBF SVM	LDA	Naive Bayes'	Decision Tree	Random Forest	AdaBoost
HILIC + CD + FTIR + UV	87.89	91.86	89.54	***98.79***	46.39	66.32	86.07	23.75
SEC + CD + FTIR + UV	92.25	**95.93**	93.96	**97.71**	51.29	72.71	92.98	28.57
HILIC + SEC + CD + FTIR + UV	90.68	93.75	93.43	***98.29***	46.14	68.04	91.00	23.82
Max accuracy for the classifier	93.96	95.96	94.54	98.82	68.86	80.64	93.54	58.43

The bold values indicate high accuracies (>95%), and the bold italic values indicate the highest accuracy for a given number of combinations of the physiochemical techniques. RBF, radial basis function.