Dissimilarity Measures for Detecting Hepatotoxicity in Clinical Trial Data*

Matthew Eric Otey Srinivasan Parthasarathy Donald C. Trost
Contact: srini@cse.ohio-state.edu

Abstract

In clinical trials, pharmaceutical companies test a new drug for the treatment of a disease by comparing the results from a large number of diseased and healthy patients exposed to either the new drug, an existing drug that treats the disease, or a placebo. The goal of these trials is to establish the safety and efficacy of the new drug. One of the primary concerns is liver toxicity, which is usually diagnosed by blood analyte tests. Often, such signals of toxicity lead to the discontinuation of drug development or withdrawal of the drug from the market. Early detection of liver toxicity can save lives and also save such companies billions of dollars. Existing approaches for detecting liver toxicity typically ignore correlations between blood analyte values which we hypothesize are essential for detecting liver toxicity. Based on this hypothesis, in this work we present novel dissimilarity measures based on principal component analysis which can be used for detecting liver toxicity and identifying subpopulations who may be susceptible. As such, our measures account for differences in the correlation structure of the data, and can be tuned by the user to account for domain knowledge. Experimental results on real clinical trial data validate our approach.

Keywords: Drug efficacy and safety analysis, outlier detection, clustering, dissimilarity measures.

1 Introduction

Drug safety issues have received an enormous amount of attention in the past year. Several large pharmaceutical companies have issued warnings or removed their drugs from the market altogether following reports of severe or deadly side-effects. Such events are harmful to the companies' public image, as well as their financial status. Each company invests large amounts of money in developing and testing new drugs. Any drug under development or clinical trials that does not make it to the market represents a huge loss for the company. Also, any drug that makes it to market but must be withdrawn represents a double loss for the company, as it is unable to recoup development costs, and may be held liable for any harmful effects of that drug. Therefore, pharmaceutical companies have intense interest in discovering any harmful effects of their drugs as early as possible, so that they can cease development or sales in order to save both lives and money.

The safety and efficacy of new drugs are determined using a set of clinical trials. Clinical trials occur in four phases. In Phase I, a new drug is tested on a relatively small group healthy subjects in order to determine the side effects and dosage levels of the drug. In Phases II and III, the drug and placebos are given to both healthy and ill subjects to determine the efficacy and safety of the drug, and to compare it to existing treatments. Phase IV studies occur after the drug has been put on the market, in order to acquire additional information on risks, benefits, and optimal dosage levels. The ability to identify harmful drugs and cease development at least one phase earlier than usual can save a pharmaceutical company billions of dollars.

In pharmaceutical clinical trials, the efficacy and safety of a compound in the treatment of a particular disease is studied by comparing the results from some healthy subjects and many patients who are randomly assigned to either the experiment compound, an existing therapies for the disease, or a placebo (for example sugar water or sugar pills often containing the inactive ingredients of the drug). The goal is to show a statistically significant improvement in two or more clinical trials relative to the control group and to show that the benefits outweigh the safety risks. Safety is studied in many ways; serial clinical laboratory blood tests are used commonly to monitor biochemical changes in the body. An organ of particular concern is the liver, which has a major detoxifying function. Abnormal blood test values related to the liver is a common reason for stopping a drug development project or causing discontinuation in a particular patient or group of patients. When liver tests are high, it is assumed that hepatotoxicity, or liver toxicity, is present. However, the rules for determining the presence of drug-induced hepatotoxicity are mostly qualitative and involve considerable clinical judgment. The current state-of-the-art in pharmaceutical research uses univariate rules applied to multiple analytes. Typically the threshold is some multiple of the upper limit of a normal range specified by the laboratory, but these rules are largely ad-hoc. More recently

*Matthew Eric Otey and Srinivasan Parthasarathy are in the Department of Computer Science and Engineering at the Ohio State University. Donald C. Trost is an employee of Pfizer, Inc. This work was sponsored by Pfizer as part of a collaborative project in pathodynamics methodology.

a rule has been employed which requires the crossing of at least two thresholds. It is known as "Hy's Rule" [4]. This is the first attempt by regulatory agencies to include two analytes in a rule for hepatotoxicity. The problems of misclassification should be obvious, because hepatotoxicity may not be so much correlated with absolute elevated blood analyte values as it is with how the analytes move together. Our hypothesis is that Hy's rule is not sufficient, and that correlations between analytes are extremely important for understanding the effects of a drug on liver toxicity.

Clinical trial data is usually in the form of a set of multivariate time series, where each variable corresponds to a blood analyte and each series corresponds to a different patient. Mining such data is particularly challenging due to factors such as unequally spaced time series, missing values, and noise due to instrumentation error and variance. Detection of hepatotoxicity requires the use of techniques that can distinguish the time series of unaffected patients and the series of hepatotoxic patients. Besides counting boundary crossings, pharmaceutical statisticians typically use univariate tests of differences in population means to quantify liver effects.

In this paper we examine the notion of quantifying the dissimilarity between different sets of data with the goal of detecting hepatotoxicity. We propose dissimilarity measures that can be used to quantify the differences between two data sets. Our hope is that our measures are more sensitive to liver toxicity than more simple techniques such as "Hy's Rule." Other applications of our measure for clinical trial data involves characterizing the differences between the different subsets of patients (for example, the differences between those on drug and those on placebo, or between males and females), and discovering subpopulations that have a greater risk of hepatotoxicity.

A suitable dissimilarity measure has several requirements. First, it must take into account as much of the information contained in the data sets as possible. For example, simply calculating the Euclidean distance between the centroids of two different data sets is ineffective, as this approach ignores the correlations present in the data sets. Second, it must be user-tunable in order to account for domain knowledge. For example, the difference in the mean analyte values for two different patients may only be determined their demographic properties (e.g. age, sex, and weight), and not by any effect of the drug. In this case, differences in the mean should be weighted less than differences in the correlations. Third, the dissimilarity measure should be tolerant of missing and noisy data, since in many domains data collection is imperfect, leading to many missing attribute values [19].

In this paper we propose the use of several dissimilarity measures based on principal component analysis (PCA). Our measures consists of components that separately take into account differences in the locations, and correlations of the data sets being compared. As such, our measure takes into account much of the information in the data set. It is also possible to weight the components differently, so one can incorporate domain knowledge into the measure. Finally, our measure is robust towards noise and missing data. We demonstrate the efficacy of the proposed measures using clinical trial data provided by Pfizer that is known to contain subjects suffering from hepatotoxicity.

The rest of the paper is organized as follows. We first briefly review related work in Section 2. We then present our dissimilarity measure in Section 3, and discuss several applications of the measure. In Section 4, we present experimental results showing the performance of our measure when used for several applications on stock market data sets. Finally in Section 5 we conclude with directions for future work.

2 Related Work

As mentioned above, there have been many metrics proposed that find the distance or similarity between the records of a data set [2, 14, 11], or the between the attributes of a data set [7, 24]. However, these metrics are defined only between a pair of records or attributes. Similarity metrics for comparing two data sets have been used in image recognition [13], and hierarchical clustering [15]. The Hausdorff distance [13] between two sets A and B is the minimum distance r such that all points in A are within distance r of some point in B, and vice-versa. Agglomerative hierarchical clustering frequently makes use of the single-link and complete-link distances between two clusters [15] to decide which pair of clusters can be merged. The single-link distance between two clusters is the minimum pairwise distance between points in cluster A, and points in cluster B, while the complete-link distance is the maximum pairwise distance between points in cluster A, and points in cluster B. There is also an average-link distance [12], which is the average of all pairwise distances between points in cluster A and points in cluster B. However, these metrics do not explicitly take into account the correlations between attributes in the data sets (or clusters). Parthasarathy and Ogihara [20] propose a similarity metric for clustering data sets based on frequent itemsets. By this metric, two data sets are considered similar if they share many frequent itemsets, and these itemsets have similar supports. This metric takes into account correlations between the attributes, but it is only applicable for data sets with categorical or discrete attributes.

There has also been work for defining distance metrics that take into account the correlations present in continuous data. The most popular metric is the Mahalanobis distance [22], which accounts for the covariances of the attributes of the data. However this can only be used to calculate the distance between two points in the same data set. Yang *et al* [25] propose an algorithm for subspace clustering (i.e. subsets of both points and attributes in a data set) that finds clusters whose attributes are positively correlated with each other. Böhm *et al* [5] modify the DBSCAN algorithm [10] by using PCA to find clusters of points that are not only density-connected, but correlation-connected as well. That is to say, they find subsets of a data set that have similar correlations. To determine if two points of the data set should be merged into a single cluster, they must be in each other's "correlation" neighborhood which is determined by a PCA-based approximation to the Mahalanobis distance. This approach is more flexible than Yang *et al*'s in that it can find clusters with negative correlations between the attributes. However, their measure is unable to find subsets of data with similar correlations that are not density-connected. Furthermore, both Yang *et al*'s and Böhm *et al* approaches are interested only in finding clusters of points within a single data set, instead of clustering multiple data sets.

Recently, Aggarwal has argued for user interaction when designing distance functions [1] between points. He presents a parametrized Minkowski distance metric and a parametrized cosine similarity metric that can be tuned for different domains. He also proposes a framework for automatically tuning the metric to work appropriately in a given domain. Based on these ideas in the next section we present a tunable metric for computing a measure of dissimilarity across data (sub)sets.

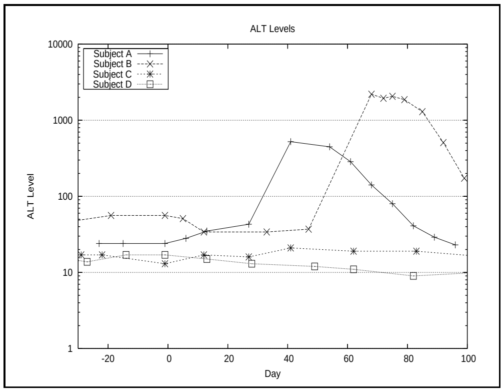

Figure 1: Plot of four different subjects' ALT levels over time.

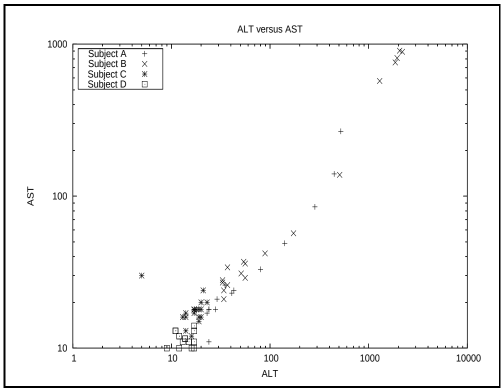

Figure 2: Plot of four different subjects' ALT levels versus their AST levels.

3 Algorithms

In this section we first discuss the challenges that mining clinical trial data sets present. We then describe a simple feature extraction procedure that enables us to represent the data alleviating the problems described. The remainder of this section is then devoted to developing our dissimilarity measure and demonstrating its utility and flexibility via a simple example, for expository simplicity.

3.1 Challenges As we discussed in Section 1, clinical trial data are presented in the form of a multivariate time series for each subject in the trial. At each time point, the values of various blood analytes are recorded. While there are many techniques for analyzing (multiple) times series data [3, 6, 9], clinical trial time series data is quite challenging. Such time series data sets suffer from irregular sampling, missing data, and varying lengths. This may be due to a variety of reasons, including missed appointments, unexplained absences, and drop outs. Furthermore, there are also several potential sources of noise. Measurement errors, laboratory bias[1], and circadian effects on analyte values (depending on when the blood sample was drawn) can be contributing factors to noise.

To illustrate these issues consider the following examples drawn from a real data source. Figure 1 shows the levels of the blood analyte ALT (alanine aminotransferase) measured over the course of the clinical trial for four different subjects. From this graph, one can see that different series have different lengths, the subjects have different numbers of samples, and on a given day, not all subjects' ALT values are sampled. However, subjects A and B suffer from hepatotoxicity, and this is marked by the large spikes in ALT values. Time series plots for other blood analytes for these subjects would show similar spikes in values.

In Figure 2 we ignore the time component and show a scatter plot of the values of the ALT analyte versus that of the AST (aspartate aminotransferase) analyte

[1] Different laboratories, where these tests are often analyzed, often have different protocols resulting in a significant variation in analyte values for the same subject.

for the same four subjects shown in Figure 1. Here we see that even if we ignore the time component, the hepatotoxic patients are obvious: There is a strong correlation between the values of the ALT and AST analytes, and these analytes both take on extreme values. Furthermore, the vector pointing in the direction of maximum variance for subject A points in nearly the same direction as that for subject B. This infers that not only the magnitude of the variance in a time series important, but its direction is as well.

3.2 Feature Extraction and Key Intuition The basis of Hy's rule, and the typical signal physicians look for when evaluating liver toxicity, is usually a significant and consistent departure from the normal levels of one or more liver analytes. Moreover, it is usually the case that not all the analytes are affected simultaneously. A conclusion one can draw from these two statements is that the covariance or in most cases the correlation among analytes should be capable of identifying such significant departures from the norm.

This key intuition leads us to the use of correlation or covariance matrices to represent patient data and subsequently the use of principal component based methods for computing dissimilarity measures for such datasets. We note that correlation and covariance matrices can easily be imputed in the presence of missing data and moreover, principal components based techniques have been shown in the literature to be noise-tolerant [19].

To summarize, our feature vector, representing the data for each subject, consists of the matrix of covariances or correlations between each pair of attributes, and the principal components derived from this matrix. The dissimilarity measures quantify the differences between the principal components of two different subjects. Additionally, we would like to note that such measures are general-purpose, and can be used to compare any two data sets, times series or not, so long has they have the same dimensionality.

3.3 Dissimilarity Measures Our goal is to quantify the dissimilarity of two k-dimensional data sets $\underline{\overline{\mathbf{X}}}$ and $\underline{\overline{\mathbf{Y}}}$. Our measures take into account the correlations between the attributes of the two data sets. In general, the dissimilarity of two data sets $\underline{\overline{\mathbf{X}}}$ and $\underline{\overline{\mathbf{Y}}}$ is denoted as $D(\underline{\overline{\mathbf{X}}}, \underline{\overline{\mathbf{Y}}})$. We define the function D in terms of two dissimilarity functions that take into account the differences in rotation, and variance between the data sets. These components are combined by means of a weighted sum, which allows one to weight the components differently, so as to incorporate domain knowledge.

The first step of using our dissimilarity measures is to the find the principal components of the data sets being compared. The principal components of a data set are the set of orthogonal vectors such that the first vector points in the direction of greatest variance in the data, the second points in the orthogonal direction of the second greatest variance in the data, and so on [17, 23]. We consider $\underline{\overline{\mathbf{X}}}$ and $\underline{\overline{\mathbf{Y}}}$ to be most similar to each other when their principal components, paired according to their ranks, are aligned and have the same magnitude, and most dissimilar when all of the components of $\underline{\overline{\mathbf{X}}}$ are orthogonal to those of $\underline{\overline{\mathbf{Y}}}$.

More formally, given a data set $\underline{\overline{\mathbf{X}}}$, consider the singular value decomposition (SVD) of its covariance matrix:

$$cov(\underline{\overline{\mathbf{X}}}) = U\Lambda_X X^T \tag{3.1}$$

where the columns of X are the principal components of the data set $\underline{\overline{\mathbf{X}}}$, arranged from left to right in order of decreasing variance in their respective directions, and Λ_X is the diagonal matrix of singular values. Note that one can also find the SVD of the correlation matrix of $\underline{\overline{\mathbf{X}}}$ as an alternative to the covariance matrix.

Having found the principal components, we can now represent each data set $\underline{\overline{\mathbf{X}}}$ as a single feature vector $F_{\underline{\overline{\mathbf{X}}}}$:

$$F_{\underline{\overline{\mathbf{X}}}} = \sqrt{\Lambda_1} \times X_1 \tag{3.2}$$

where X_1 is the first principal component of the data set, or the first column of X in Equation 3.1, and Λ_1 is its corresponding eigenvalue. That is to say each data set is represented by the scaled primary principal component *vector* pointing in the direction of greatest variance.

Having such a feature vector, we can then apply any standard distance metric. For example, applying the Euclidean distance metric:

$$D_e(F_{\underline{\overline{\mathbf{X}}}}, F_{\underline{\overline{\mathbf{Y}}}}) = |F_{\underline{\overline{\mathbf{X}}}} - F_{\underline{\overline{\mathbf{Y}}}}|_2 \tag{3.3}$$

on the first principal component derived from the covariance matrix of the data would result in a value that simultaneously measures differences in direction and magnitude of the vector. We have also developed an alternative distance metric called the Projection distance:

$$D_p(F_{\underline{\overline{\mathbf{X}}}}, F_{\underline{\overline{\mathbf{Y}}}}) = 1 - \frac{F_{\underline{\overline{\mathbf{X}}}} \cdot F_{\underline{\overline{\mathbf{Y}}}}}{max(|F_{\underline{\overline{\mathbf{X}}}}|^2, |F_{\underline{\overline{\mathbf{Y}}}}|^2)} \tag{3.4}$$

which measures the length of the projection of the shorter vector onto the longer one.

These two measures can be extended to account for the differences in the mean of the data sets. First we define the dissimilarity of the means of the data sets as follows:

$$D_\mu(\underline{\overline{\mathbf{X}}}, \underline{\overline{\mathbf{Y}}}) = |\mu_{\underline{\overline{\mathbf{X}}}} - \mu_{\underline{\overline{\mathbf{Y}}}}|_2. \tag{3.5}$$

that is to say, the Euclidean distance between the centroids of the two data sets. We can then define the extended D_e measure as follows:

$$D_e(\overline{\underline{\mathbf{X}}}, \overline{\underline{\mathbf{Y}}}) = \beta_0 + \beta_1 \times D_\mu + \beta_2 \times D_e \tag{3.6}$$

and the extended D_p measure as:

$$D_p(\overline{\underline{\mathbf{X}}}, \overline{\underline{\mathbf{Y}}}) = \beta_0 + \beta_1 \times D_\mu + \beta_2 \times D_p. \tag{3.7}$$

This formulation allows us to weight differences in the means and correlations according to domain information. For example, in clinical trial data, differences in the means of the observations of two different subjects may be caused more by differences in demographic characteristics (e.g. sex, age, weight) than by any effect of the drug, and so one would want to weight the differences in correlations higher.

Finally, we note that we can generalize these measures to account for all the principal components as follows. Let $F^i_{\underline{\mathbf{X}}}$ be the feature vector for the ith component:

$$F^i_{\underline{\mathbf{X}}} = \sqrt{\Lambda_i} \times X_i. \tag{3.8}$$

Then the D_e measure can be generalized as:

$$D'_e(\overline{\underline{\mathbf{X}}}, \overline{\underline{\mathbf{Y}}}) = \sum_{i=1}^{k} D_e(F^i_{\underline{\mathbf{X}}}, F^i_{\underline{\mathbf{Y}}}), \tag{3.9}$$

while the D_p measure has a similar general form:

$$D'_p(\overline{\underline{\mathbf{X}}}, \overline{\underline{\mathbf{Y}}}) = \sum_{i=1}^{k} D_p(F^i_{\underline{\mathbf{X}}}, F^i_{\underline{\mathbf{Y}}}). \tag{3.10}$$

3.3.1 Missing Data Our measure is also robust to missing data, which commonly occurs in clinical trial data. Reasons for missing data include the fact that subjects may not show up at appointments; the protocol may not require a complete set of tests; or a blood sample may be mishandled or may contain interfering ingested substances. If a data set $\overline{\underline{\mathbf{X}}}$ has records with missing attribute values, and assuming that the data has a normal distribution, one can use the Expectation-Maximization algorithm [8] to find the maximum-likelihood values of the centroid $\mu_{\underline{\mathbf{X}}}$ and the covariance matrix $cov(\overline{\underline{\mathbf{X}}})$. The principal components one finds are the sample principal components [16], and one can develop confidence intervals to test the closeness to the true (population) principal components. If the missing data is not excessive, then the maximum likelihood/sample estimates of the components will be accurate, and the computation of the dissimilarity metric can continue as before. Other approaches for handling missing data involve just ignoring records with missing data completely.

	A	B	C	D
A	—	751	271	284
B	751	—	930	936
C	271	930	—	29.5
D	284	936	29.5	—

Table 1: Euclidean dissimilarity (D_e).

	A	B	C	D
A	—	0.779	0.974	1.024
B	0.779	—	0.995	1.002
C	0.974	0.995	—	1.396
D	1.024	1.002	1.396	—

Table 2: Projection dissimilarity (D_p).

3.4 Example Here we present an example to illustrate how each dissimilarity performs with respect to the clinical trials subjects presented in Figures 1 and 2. We examine the basic D_e and D_p measures as defined in equations 3.3 and 3.4. For this example, we use the values of the eight primary blood analytes responsible for measuring liver function. These serum analytes are ALT, AST, GGT (γ-glutamyltransferase), LD (lactate dehydrogenase), ALP (alkaline phosphatase), total bilirubin, total protein, and albumin.

In Table 1 we present the pairwise dissimilarities for the D_e measure. For this example we perform SVD using the covariance matrix as opposed to the correlation matrix (see Equation 3.1) As can be seen, the two non-hepatotoxic subjects (C and D) are the most similar, followed by the hepatotoxic patients A and B. This result is visualized as a dendrogram in Figure 3.

We next look at the D_p measure. The results are presented in Table 2. Unlike D_e, the D_p measure identifies the two hepatotoxic subjects (A and B) as being the most similar, followed by subjects C and D. The resulting dendrogram is shown in Figure 4. This

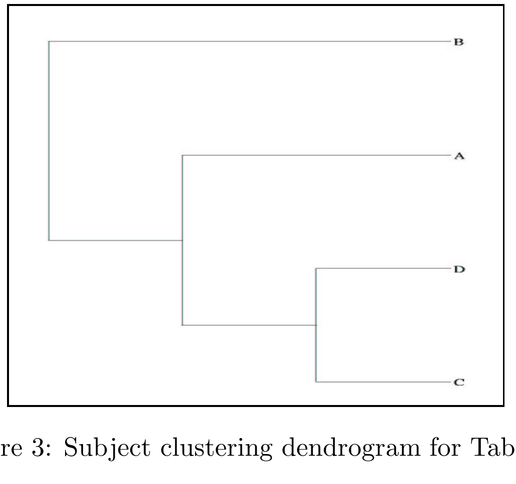

Figure 3: Subject clustering dendrogram for Table 1.

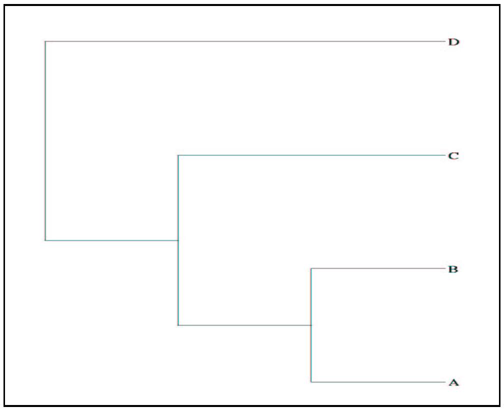

Figure 4: Subject clustering dendrogram for Table 2.

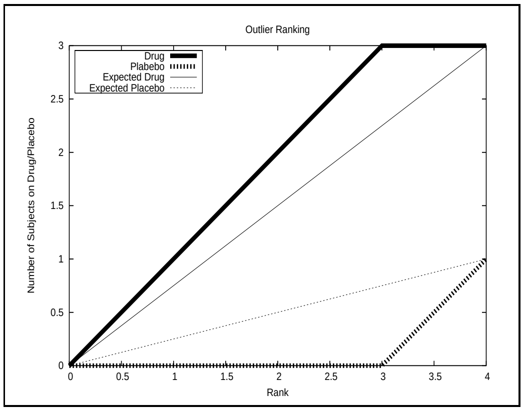

Figure 5: Outlier ranking using the Euclidean dissimilarity measure (D_e).

result is to be expected, as the hepatotoxic patients have principle component vectors of nearly the same length pointing in nearly the same direction. This example illustrates the differences between these two measures: The D_e measure tends to be more sensitive to the magnitude of the feature vectors, while the D_p measure tends to be more sensitive to the direction of the feature vectors.

3.5 Applications In this section we present an overview of how our dissimilarity measures can be used to analyze the clinical trial data. The techniques we consider are anomaly (outlier) detection, and data set clustering.

3.5.1 Anomaly Detection Detection of anomalies or outliers in clinical trial data is very important. Subjects' analyte values may be anomalous for many reasons related to sample processing including subject ingestion of interfering substances, sampling handling conditions, analyzer error, and transcription error. If these data points can be identified and the cause attributed to a non-treatment-related event, then the data point may need to be removed from a particular analysis. Subjects' values may be anomalous because they are having abnormal reactions to the drug. If this is the case, the drug maker may want to study more subjects similar to the anomalous ones to see if they are true anomalies or indicative a small sub-populations that may have toxic reactions to the drug.

Using our dissimilarity measures, it is straightforward to implement basic outlier detection algorithms such as those described in [18]. These are nested-loop approaches that calculate the dissimilarity between each pair data points (or in our case, each pair of subjects). Having calculated these values one can rank the data points (subjects) according to several different approaches. The KTH approach ranks the subjects according to their dissimilarity from their kth most similar subject. The KSUM approach ranks the subjects according to the sum of the dissimilarities from the k most similar subjects. Finally, the NSUM approach ranks the subjects according to the dissimilarities from all other $(n-1)$ subjects. In this paper we use the KSUM approach.

Once we have these outlier rankings for all the subjects in a given study, we can use them to determine not only which subjects are the most anomalous, but also to determine if the drug being studied has any appreciable effect. For example, if we examine a ranking of the subjects, we would expect the hepatotoxic patients to be highest-ranked, followed by the remaining subjects who were on the drug, and finally the patients who were given a placebo. However, a drug that has little or no effect on the liver tests is less likely to cause hepatotoxicity, and subjects on such a drug should not be very dissimilar from those on placebo, meaning that the ranking would be random.

To examine the effects of the drug being studied, we use graphs such as the one in Figure 5. In this graph, we plot the cumulative number of subjects on drug and on placebo given the outlier ranking using thick lines. The thin lines express the expected cumulative number of subjects on drug or placebo for a given ranking assuming the ranking is random. Of the example subjects presented in Section 3.1, A, B, and D were taking the drug, while C was on placebo. If we use the NSUM approach on Euclidean distance measure results in Table 1, we would rank B as the most anomalous subject, followed by A, D, and C. This is an "optimal" ranking, in that it ranks the subject on placebo last, and the hepatotoxic patients first. The graph arising from this ranking is presented in Figure 5.

3.5.2 Clustering The dissimilarity measures we present above allow us to easily perform clustering of the subjects, as we did in the example in Section 3.4

(see the dendrograms in Figures 3 and 4). Finding clusters of subjects in clinical trial data is helpful in that it allows us to identify sub-populations who may have a greater risk of hepatotoxicity, sub-populations on whom the drug may have little or no effect, sub-populations that may have a higher risk of severe side-effects, et cetera. This allows the drug makers to determine the efficacy of the drug, to determine dosage levels for different patients, and to determine if the side-effects are too severe or widespread to continue development of the drug.

It is straightforward to perform agglomerative hierarchical clustering of data sets using our dissimilarity measures. If one has n data sets, one can construct an n by n table containing the pairwise dissimilarities of the data sets. Once this table has been constructed, one can use any distance metric (e.g. single-link or complete-link) to perform the hierarchical clustering. We present experimental results on using hierarchical clustering for the clinical trial data in Section 4.3. This table also facilitates non-hierarchical clustering approaches, such as the k-medoid approach [12]. This works by selecting several data sets at random to be medoids of the clusters, and then assigning the remaining data sets to a cluster with the most similar medoid. After this phase, the medoids are checked to see if replacing any of them with other data sets would reduce the dissimilarity in their respective clusters. If so, the process repeats until no medoids are replaced, or some other stopping criterion is met.

4 Experimental Results

In this section we evaluate the efficacy of the proposed approach on clinical trials data obtained from Pfizer, Inc. The end objective of this study is to evaluate the impact of drug on liver analytes in order to understand the hepatotoxicity effects of the drug. Below we describe details about the datasets used in this evaluation.

4.1 Setup The first dataset we use, henceforth referred to as $D1$, consist of a set of subjects suffering from diabetes, who were given either a placebo (a formulation that includes only the inactive ingredients) or the drug under study (drug A). Since we are primarily concerned with hepatotoxicity, under suggestions from our domain experts we only considered data from eight serum analytes (often referred to in the literature as the liver panel or liver function tests): ALT, AST, GGT, LD, ALP, total bilirubin, total protein, and albumin. Using advice from a domain expert, we used the logarithm transformation of the first six analytes' values (total protein and albumin are excepted), unless otherwise noted. This dataset consisted of 446 patients on placebo and 680 patients on drug. This drug was under development but development was discontinued in Phase III for various reasons including possible hepatotoxicity.

The second dataset we use, henceforth referred to as $D2$, consisted of a set of post-menopausal women, who again were given either a placebo or one of two drugs (B, C) (both are different from the one used in $D1$). Again, we limited our focus to the liver panel. This dataset consisted of 201 patients on placebo, 41 patients on drug B, and 126 patients on drug C. All three of these drugs were marketed drugs and were expected to have little or no hepatotoxicity.

Both datasets suffer from the problems we mentioned earlier. They contain missing data, unequally spaced time series data for different patients, some patients had many readings over a period of time, others had much fewer etc. As noted earlier we transformed the data from each patient into a feature vector as described earlier in Section 3.3. Since the differences in the mean are not significant in these data sets, we use the basic forms of the D_e and D_p measures defined in equations 3.3 and 3.4 in these experiments. All of our implementations are done using Octave, an open-source version of Matlab. All dendrograms were visualized using the njplot software [21].

4.2 Anomaly Detection In our first experiment, we want to see how our dissimilarity measures perform on the clinical trial data set of diabetic patients. As noted earlier we have two groups of patients: one on placebo, and another on the drug under study. The experiment we conduct is to flag outliers from the dataset using the dissimilarity measures discussed in the previous section. The null hypothesis is that if the drug does not result in hepatotoxicity, then the outliers are likely to be flagged at random from each group. Note that previous to drug intake the distributions of the two groups are nearly identical. A significant deviation from random, or to be more exact, if the people on drug tend to be flagged as outliers with a greater probability than expected, then a reasonable conclusion would be that there may be a hepatotoxic effect resulting from drug intake.

In Figure 6 we plot the outlier ranking arising from both the D_e and D_p measures for the top 10% (113) of the outliers in this dataset. We observe that in both cases the expected number of outliers from the drug group is significantly exceeded by the actual number indicating a clear signal that the drug under question is causing a change in analyte behavior in the patients being flagged as outliers. The D_e measure appears to be a little more sensitive to this phenomenon than the D_p measure. We would like to note that Phase III continued for approximately two more years after these cases were completed. Had this signal been detected at that time, Pfizer might have been able to save on the resources it expended to continue Phase III.

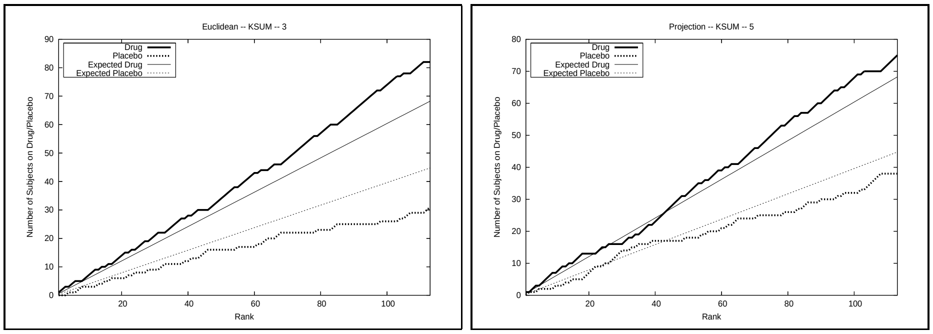

Figure 6: Top outlier rankings for $D1$ using the (A) D_e, and (B) D_p dissimilarity measures.

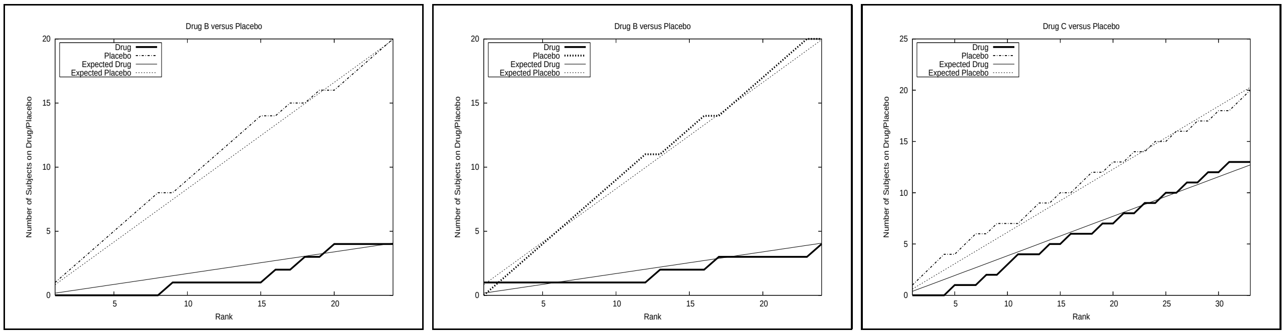

Figure 7: Top Outlier rankings for (A) using D_e on Drug B in $D2$, (B) using D_p on Drug B in $D2$, and (C) using D_e on Drug C in $D2$

In our second experiment we evaluate the performance of our method on the second dataset composed of healthy post-menopausal women. As these are healthy women taking either a placebo or drugs on the market with no known hepatotoxic effects, we expect there to be few (if any) true outliers. In such a case, our plots should show that the number of outliers corresponding to subjects on drug should be at or below the expected levels. In Figure 7(A) and (B), we plot the top 10% of the outliers for both drugs using the D_e measure. As we hypothesize, the mixture of subjects on drug and placebo marked as outliers are near the expected levels. In Figure 7(C), we show the plot resulting from using the D_p measure on Drug B. Again, the mixture of subjects on drug and placebo is near the expected level. The plot when D_p is run on Drug C is similar and not included here. As can be seen, there is little or no difference from the expected numbers for placebo and drug, inferring that there are no subjects suffering from hepatotoxicity in this case which is what we would expect from two drugs that are currently on the market. Both measures are equally effective in this experiment and there is little to choose among them.

In our third experiment we examined what effect varying the number of principal components has on the outlier rankings. In this case, we varied the number of components used by D'_e (see Equation 3.9) between 2 and 4 and applied it to the data sets for Drug A. The results can be seen in Figure 8. For reference, recall that Figure 6(A) shows the D'_e measure with only 1 component. As can be seen from the graphs, when we move from 1 component to 2, there is little change. However, when we move to 3 components, we mark significantly more subjects on drug as outliers. This appears to be the optimum number of components in this case, for when we move to 4 components, the sensitivity decreases somewhat. Though they are not pictured here, the results for D'_p are similar.

These experiments demonstrate and advantage of our approach over Hy's rule. They show that we are capable on not only finding important differences in magnitude, but also in direction (correlation) that may be missed by Hy's rule.

4.3 Data Set Clustering In our final experiment we demonstrate the utility of using our dissimilarity measures to perform clustering. In Figure 9 we present the dendrogram resulting from performing complete-link hierarchical clustering. In this case we use a subset of the subjects corresponding to all males with diabetes who were taking the drug being studied, for a total of 450 subjects. We use the Euclidean dissimilarity measure D_e and the covariance matrices. We find that clustering results in an intuitive grouping of the subjects. For example, we look at two different branches of the dendrogram in Figures 10 and 11.

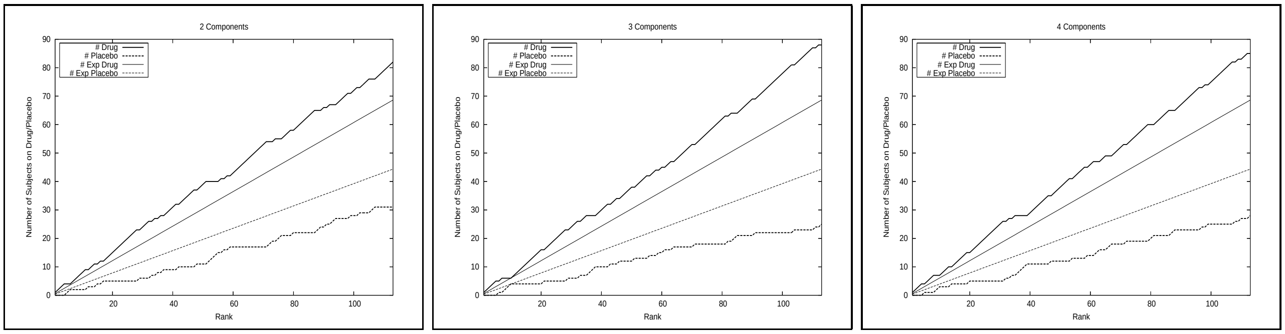

Figure 8: Top Outlier rankings for Drug A as function of the number of components used in D'_e: (A) 2 components, (B) 3 components, (C) 4 components.

In Figure 10, we examine the top-most branches of the dendrogram in Figure 9. These branches correspond to a cluster of subjects with relative low spikes in analyte values. These spikes may not be large enough to be considered a sign of hepatotoxicity according to Hy's rule. In Figure 11 we examine the bottom-most branches of the full dendrogram, and find that these correspond to a cluster of subjects with very large spikes in analyte values, nearly an order of magnitude larger than those in Figure 10. Although we only plot the ALT levels here, we note these spikes extend to the other blood analytes as well and affect the overall covariances. Hy's rule would definitely categorize the cases in Figure 11 as being hepatotoxic, whereas the cases in Figure 10 may or may not be categorized as hepatotoxic depending on the amplitude of the spike. Other branches show different behaviors that may not be indicative of hepatotoxicity, but may be related to the subjects' demographic or other health attributes, which may aid in determining dosage levels.

5 Conclusion

Efficient and precise analysis of clinical trial data is very important to pharmaceutical companies, as it allows them to determine the efficacy and safety of a drug. Pharmaceutical companies want to halt development on unsafe and ineffectual drugs as early as possible in order to save on development costs and to avoid unnecessary complications and severe side-effects that may lead to liability suits if the drug were to reach the market. Current approaches for detecting hepatotoxicity in clinical trial data sets have limited effectiveness, due to the fact that they typically ignore correlations between blood analytes. Since clinical trial data is the form of irregular time series, it is difficult to apply standard statistical approaches to determine the (dis)similarity of two or more subjects. In this paper we presented several dissimilarity measures for data sets that takes into account the means and covariance structures of the data sets. Our results on real clinical trial data show that our measures can be very helpful in detecting true hepatotoxicity and finding subpopulations of subjects who may have different reactions to the drug under study.

References

[1] Charu C. Aggarwal. Towards systematic design of distance functions for data mining applications. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 9–19, August 2003.

[2] Rakesh Agrawal, Christos Faloutsos, and Arun N. Swami. Efficient similarity search in sequence databases. In *FODO '93: Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*, pages 69–84. Springer-Verlag, 1993.

[3] Francesco Audrino and Peter Buhlmann. Synchronizing multivariate financial time series. Technical Report Research Report 97, Seminar fur Statistik, May 2001.

[4] Einar Bjornsson and Rolf Olsson. Outcome and prognostic markers in severe drug-induced liver disease. *Hepatology*, 42(2):481–489, August 2005.

[5] Christian Bohm, Karin Kailing, Peer Kroger, and Arthur Zimek. Computing clusters of correlation connected objects. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 455–466. ACM Press, 2004.

[6] Tak chung Fu, Fu lai Chung, Robert Luk, and Chak man Ng. Financial time series indexing based on low resolution clustering. In *Proceedings of the IEEE International Conference on Data Mining Workshop on Temporal Data Mining: Algorithms, Theory and Applications*, November 2004.

[7] Gautam Das, Heikki Mannila, and Pirjo Ronkainen. Similarity of attributes by external probes. In *Knowledge Discovery and Data Mining*, pages 23–29, 1998.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B39:1–38, 1977.

[9] Emre Erdogan, Sheng Ma, Alina Beygelzimer, and Irina Rish. Statistical models for unequally spced time series. In *SIAM*, 2004.

[10] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama Fayyad, editors, *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland, Oregon, 1996. AAAI Press.

[11] Roy Goldman, Narayanan Shivakumar, Suresh Venkatasubramanian, and Hector Garcia-Molina. Proximity search in databases. In *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, pages 26–37, 24–27 1998.

[12] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.

[13] D. P. Huttenlocher, G. A. Klanderman, and W. A. Rucklidge. Comparing images using the hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(9):850–863, 1993.

[14] H. V. Jagadish, Alberto O. Mendelzon, and Tova Milo. Similarity-based queries. In *PODS '95: Proceedings of the fourteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 36–45, New York, NY, USA, 1995. ACM Press.

[15] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[16] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, fifth edition, 2002.

[17] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[18] E. Knorr and R. Ng. A unified notion of outliers: Properties and computation. In *ACM SIGKDD*, 1997.

[19] Srinivasan Parthasarathy and C. C. Aggarwal. On the use of conceptual reconstruction for mining massively incomplete data sets. *IEEE Transactions on Knowledge and Data Engineering*.

[20] Srinivasan Parthasarathy and Mitsunori Ogihara. Clustering distributed homogeneous datasets. In *PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 566–574, London, UK, 2000. Springer-Verlag.

[21] G. Perriere and M. Gouy. Www-query: An online retrieval system for biological sequence banks. *Biochimie*, (78):364–369.

[22] Jose C. Principe, Neil R. Euliano, and W. Curt Lefebvre. *Neural and Adaptive Systems: Fundamentals through Simulations*. John Wiley and Sons, 2000.

[23] Richard Reyment and K. G. Joreskog. *Applied Factor Analysis in the Natural Sciences*. Cambridge University Press, 1996.

[24] R. Subramonian. Defining diff as a data mining primitive. In *KDD 1998*, 1998.

[25] Jiong Yang, Wei Wang, Haixun Wang, and Philip Yu. δ-clusters: Capturing subspace correlation in a large data set. In *Proceedings of the 18th International Conference on Data Engineering (ICDE'02)*, page 517. IEEE Computer Society, 2002.

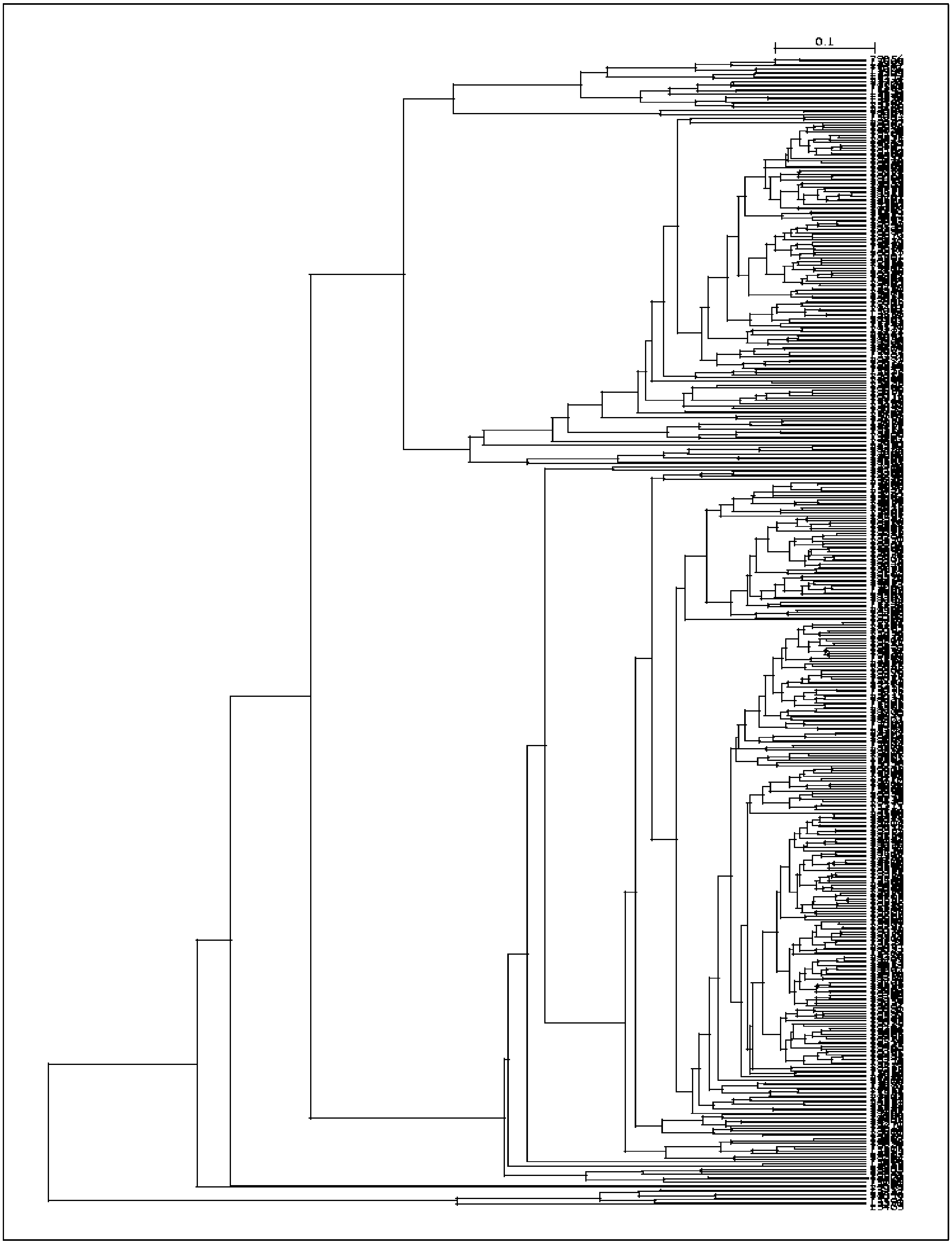

Figure 9: Dendrogram resulting from clustering males with diabetic neuropathy taking drug.

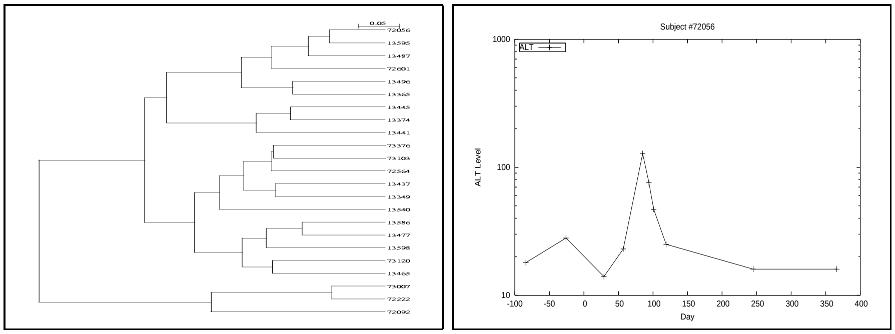

Figure 10: (A) Upper branch of complete dendrogram containing subjects with (B) small spikes in analyte values.

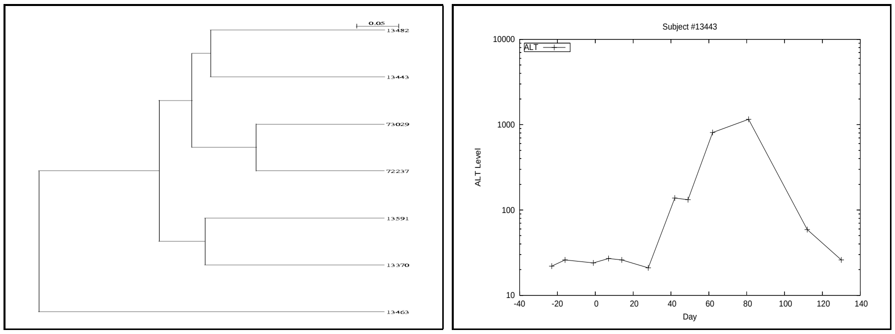

Figure 11: (A) Lower branch of complete dendrogram containing subjects with (B) large spikes in analyte values.